Enhanced Mobile Computing Using Cloud Resources

Andrew James Paverd

Submitted to the Department of Electrical Engineering,
in partial fulfilment of the requirements for the degree of

Master of Science in Engineering

at the

University of Cape Town

December 2011

Supervisor:	Prof. M. R. Inggs
Co-supervisor:	Dr S. L. Winberg

Declaration

I know the meaning of plagiarism and declare that all the work in this document, save for that which is properly acknowledged, is my own. This minor dissertation is being submitted in partial fulfilment of the requirements for the degree of Master of Science in Engineering at the University of Cape Town. This minor dissertation has not been submitted before for any degree or examination at any other university.

Signed by candidate

Signature Removed

Andrew James Paverd

1 December 2011

Abstract

The purpose of this research is to investigate, review and analyse the use of cloud resources for the enhancement of mobile computing. Mobile cloud computing refers to a distributed computing relationship between a resource-constrained mobile device and a remote high-capacity cloud resource. Investigation of prevailing trends has shown that this will be a key technology in the development of future mobile computing systems. This research presents a theoretical analysis framework for mobile cloud computing. This analysis framework is a structured consolidation of the salient considerations identified in recent scientific literature and commercial endeavours. The use of this framework in the analysis of various mobile application domains has elucidated several significant benefits of mobile cloud computing including increases in system performance and efficiency. Based on recent scientific literature and commercial endeavours, various implementation approaches for mobile cloud computing have been identified, categorized and analysed according to their architectural characteristics. This has resulted in a set of advantages and disadvantages for each category of system architecture. Overall, through the development and application of the new analysis framework, this work provides a consolidated review and structured critical analysis of the current research and developments in the field of mobile cloud computing.

Acknowledgements

This research endeavour has been made possible through an Innovation Masters Scholarship from the Department of Science and Technology (DST) and the National Research Foundation (NRF) of South Africa. Additional funding has been provided through a Postgraduate Merit Scholarship from the University of Cape Town. I would like to express my gratitude to these organizations for their financial support of this research.

I would also like to express my gratitude to the following people and groups who have assisted me during this research endeavour:

Professor Michael Inggs for the invaluable guidance, supervision and support he has provided throughout this process.

Dr Simon Winberg for the technical input and guidance he has provided in his capacity as co-supervisor.

Professor Barry Downing and the Department of Electrical Engineering for facilitating my postgraduate study at the University of Cape Town.

My colleagues at the university and, in particular, the members of the Radar Remote Sensing Group for the stimulating academic environment they have provided.

My family and friends for their unfailing support throughout this process.

Contents

List of Figures

List of Tables

Nomenclature

BOINC	—	Berkeley Open Infrastructure for Network Computing
CBIR	—	Content-Based Image Retrieval
GB	—	Gigabyte
GPS	—	Global Positioning System
GPU	—	Graphics Processing Unit
HA	—	High Availability
HCI	—	Human-Computer Interaction
HIPAA	—	Health Insurance Portability and Accountability Act
HPC	—	High Performance Computing
HTTP	—	Hypertext Transfer Protocol
IaaS	—	Infrastructure as a Service
ICT	—	Information and Communication Technology
IMS	—	Internet Protocol Multimedia Subsystem
IP	—	Internet Protocol
ITU	—	International Telecommunication Union
LAN	—	Local Area Network
LBS	—	Location-Based Service
M2M	—	Machine-to-Machine

MAC	—	Message Authentication Code
MAN	—	Metropolitan Area Network
Mbps	—	Megabits per second
MCCD	—	Mobile Computing & Communication Device
MNO	—	Mobile Network Operator
NFC	—	Near Field Communication
NIST	—	National Institute of Standards and Technology
OGSI	—	Open Grid Services Infrastructure
OS	—	Operating System
PaaS	—	Platform as a Service
PAN	—	Personal Area Network
PC	—	Personal Computer
PDA	—	Personal Digital Assistant
PHI	—	Protected Health Information
QoE	—	Quality of Experience
RAM	—	Random Access Memory
RAT	—	Radio Access Technology
SaaS	—	Software as a Service
SDK	—	Software Development Kit
SLA	—	Service Level Agreement
UI	—	User Interface
UMTS	—	Universal Mobile Telecommunications System
WAN	—	Wide Area Network
WLAN	—	Wireless Local Area Network

Chapter 1

Introduction and Overview

1.1 Introduction

Mobile Computing & Communication Devices (MCCDs) such as cellular phones, smartphones and tablet computers are becoming increasingly prevalent. Mobile computing is a paradigm in which these mobile devices are used for computational purposes. In this paradigm, the computational device can be used whilst mobile. There is a growing trend towards the use of mobile computing powered by modern MCCDs. However, due to the fundamental constraints of mobility, mobile devices are inherently resource-constrained relative to non-mobile systems in terms of their computational capabilities. It has been proposed in recent scientific literature that the emerging cloud computing paradigm, which involves the use of remote high-capacity elastic computational resources, could be used to augment the capabilities of mobile devices. Using this paradigm, MCCDs could benefit from on-demand access to the computational capacity provided by cloud resources via a network such as the internet. These cloud resources represent a dynamic combination of data processing, storage and communication capacity which is significantly greater than that of a mobile device. This distributed computing relationship is known as *mobile cloud computing* [1] [2] [3].

The overall aim of this work is to provide a consolidated review and structured critical analysis of current research and developments within the field of mobile cloud computing. This introductory chapter presents background information, defines the scope and objectives of this project and outlines the structure of this document. **Section 1.2** provides relevant background information and **Section 1.3** defines the two core concepts in the context of this work. Based on these definitions, the scope and objectives of this project are outlined in **Section 1.4** and the research methodology is explained in **Section 1.6**. The overall structure of this minor dissertation is presented in **Section 1.5**. The main conclusions and outcomes of this research endeavour are summarized in **Section 1.7**.

1.2 Background

Through recent advances in both mobile hardware and software technology, modern MCCDs provide a variety of capabilities over and above those of basic communication devices. MCCDs generally feature significant computational capabilities, advanced user interfaces and a variety of sensors and peripherals. According to data from the International Data Corporation (IDC), there were over 1.3 billion shipments of mobile phones worldwide in the year 2010 [4]. By comparison, there were 346 million Personal Computers (PCs) shipped worldwide over the same period [5].

A definite trend towards mobile internet connectivity has been observed. Worldwide data from the International Telecommunication Union (ITU) shows that the number of mobile broadband users has exceeded the number of fixed broadband users since 2008 [6]. It is estimated that in 2011 there were 17.0 mobile broadband connections per 100 inhabitants worldwide [6].

The computational capacity of MCCDs is used to run mobile software applications, commonly referred to as *mobile apps*. Mobile apps extend the functionality of the mobile device. There is already a broad range of available apps as indicated by data from the IDC showing that over 300,000 different mobile apps were developed during the period 2008 to 2010 [7]. It is common for these apps to be made available to users through central online repositories known as *app stores*. Gartner predicts the mobile app sector will grow to more than ten times its current size and generate revenues of approximately USD 58 billion by 2014 [8]. Since the use of mobile apps is a form of mobile computing, the growing popularity of these apps indicates a strong trend towards mobile computing.

The cloud computing paradigm is currently experiencing a period of rapid technological innovation and is gaining momentum in various areas of computing. By using high-capacity infrastructure such as purpose-built data centres, clouds can provide a combination of data processing and storage capacity which is significantly greater than that of a mobile device. Another key characteristic of the cloud paradigm is that its highly elastic nature allows rapid provisioning of computational resources. This elasticity is achieved through the use of virtualization technology as explained in **Section 2.4**.

Since cloud computing is still an emerging technology, the terminology for describing different aspects of this concept is currently in a state of flux. However, there are ongoing efforts to achieve standardization on specific aspects of the relevant terminology. The *"NIST Definition of Cloud Computing"* is one such effort which presents recommendations from the National Institute of Standards and Technology (NIST) [9].

According to the NIST definition, cloud computing is composed of the following three service models:

- Cloud Software as a Service (SaaS)
- Cloud Platform as a Service (PaaS)
- Cloud Infrastructure as a Service (IaaS)

The NIST definition also lists the following four deployment models for cloud infrastructure:

- Private clouds
- Community clouds
- Public clouds
- Hybrid clouds

These service and deployment models mostly coincide with recent general literature on the topic of cloud computing. Authors of scientific literature have also accepted these definitions to a certain extent. For example, in their seminal paper, *"Above the Clouds: A Berkeley View of Cloud Computing"*, Armbrust et al. agree with the NIST deployment models but avoid the service model terminology because they feel it lacks the required clarity to distinguish between the various service models [10].

Recently, it has been proposed that cloud resources could be used to enhance the computational capabilities of mobile devices. This paradigm is referred to as *mobile cloud computing* and has been discussed by Kumar and Lu [1], Chun et al. [2] and Marinelli [3] amongst others. Chun et al. explain how mobile computing is currently being used in a wide variety of applications and claim that *"Mobile cloud computing is the next big thing"* [2].

Mobile cloud computing is currently going through a critical phase of its development since this new paradigm is gaining momentum in both academic research as well as commercial endeavours. This means that there is an increasing amount of information available on this topic. In order to continue research and development in this field, it is necessary to consolidate the existing information and represent it in a structured form. Simoens et al. have reviewed and analysed a number of mobile cloud computing systems specifically in the area of remote graphical rendering [11]. They have proposed that future research should focus on *"the design of an overall framework"* for mobile cloud computing [11]. This motivates one of the key objectives of this work.

For the purposes of this work, definitions of key terms are established in

Section 1.3. The scope and objectives of this research endeavour are presented in **Section 1.4**.

1.3 Definitions

This section defines the terms *'Mobile Computing & Communication Device'* and *'Cloud Computing'* specifically in the context of this work. These definitions are important not only for clarity but also because they affect the scope of this research. Since these are both areas of technology in which rapid development is currently taking place, it is important that these definitions are not so specific as to be made obsolete by the next iteration in the technology development cycle. In order to achieve this, these two concepts are defined in terms of their functional characteristics. Although specific examples are discussed throughout this work, these are used only for explanatory purposes. The concepts presented in this research are applicable to all systems which comply with the following definitions:

1.3.1 Mobile Computing & Communication Device

A Mobile Computing & Communication Device (MCCD) is any device which is designed for mobility and which provides some degree of computational capacity and network connectivity whilst mobile. In this context, the concept of mobility differs from that of portability. A portable device can be operated whilst stationary in different geographical locations. A mobile device is portable but can also be used whilst moving between different locations.

A device specifically designed for mobility inherently exhibits two important characteristics. Firstly, the device has a finite energy source that stores the energy required for mobile operation. At present, most mobile devices use rechargeable batteries. Secondly, the design of a mobile device ensures portability by limiting the physical mass and dimensions of the device. It is not possible to define mobile devices in terms of absolute physical size. This is illustrated by devices which are based on the tablet form-factor and are significantly larger than smartphones but are still classified as mobile devices.

Computational capacity means that the mobile device has the capability to perform useful computational operations whilst mobile. Most modern MCCDs run a mobile Operating System (OS) which allows the execution of various software applications. Network connectivity must be provided by some form of wireless communication network which allows the device to maintain a connection whilst mobile. This is usually provided by cellular communication networks or Wireless Local Area Networks (WLANs).

In general literature, the many devices included in this definition are often categorized using terms such as *'feature phones'*, *'smartphones'*, or even *'superphones'*. However, from a technical perspective, these distinctions do not affect the scope of this definition or the subsequent research. In this document, the terms *'mobile device'* or *'device'* refer to an MCCD.

1.3.2 Cloud Computing

Cloud computing involves the use of remote high-capacity elastic computational resources which are accessed via a network such as the internet. These resources must provide significantly greater computational capacity than is available on a mobile device. Exact quantification of this capacity is deliberately avoided due to the rapid pace of development of cloud computing technology. For the purposes of this work, the computational capacity of a cloud resource is assumed to be at least an order of magnitude greater than that of a mobile device.

Cloud resources can be provisioned on-demand in a highly elastic manner. There are various services which could be provided through this paradigm but for the purpose of this research, these can be abstracted as a dynamic combination of data processing, data storage and network connectivity. Most current applications based on this paradigm demonstrate these capabilities by transferring some quantity of data into the cloud, performing computational operations on the data and storing the results. Since cloud resources are accessed via a network interface, it is assumed that these resources are also capable of communicating with other nodes on the network.

Although there is often a usage-based or time-based financial cost associated with the use of public cloud infrastructure, this cost does not form part of the definition for this work. Whilst the importance of this financial cost is not disputed, the reason for its exclusion is that, from a technical perspective, it is equally possible to use an alternative deployment model such as private or hybrid cloud infrastructure. These alternative approaches are not subject to the same financial cost structure.

1.4 Scope and Objectives

The overall aim of this work is to provide a consolidated review and structured critical analysis of current research and developments in the field of mobile cloud computing. In order to achieve this, four major objectives for this minor dissertation are defined as follows:

- The first objective is to investigate the current situation and prevailing trends in mobile computing. This includes analysing how mobile computing can be enhanced to meet the increasing computational requirements of new mobile applications. This serves to contextualize the work and provides insight into the fundamental motivation for enhanced mobile computing.

- Various important considerations relevant to mobile cloud computing have been discussed in recent scientific literature. It is often the case that a particular research effort addresses an incomplete subset of these points when describing a system based on this technology. The second objective is to identify and consolidate these various considerations and to define a structured theoretical analysis framework for mobile cloud computing.

- There currently exists a broad range of application domains for mobile computing. Some of these domains have been proposed as good candidates for mobile cloud computing and others are already making use of this technology. The third objective of this work is to analyse and compare these mobile application domains using the defined analysis framework. This leads to a set of conclusions for each application domain in the context of mobile cloud computing.

- In terms of the actual implementation of mobile cloud computing, some strategies and approaches have been proposed through both recent scientific literature and commercial endeavours. Although these differ in complexity, they often exhibit similarities based on their functional characteristics and so can be categorized along these lines. Building on the conclusions of the preceding objectives, the fourth objective is to categorize, analyse and compare these implementation approaches by drawing on certain aspects from the analysis framework.

In this work, the term *'costs'* is used in its most fundamental form to refer to the use or consumption of finite resources. For example, in the context of mobile devices, energy costs refer to the consumption of energy from the device's battery and communication costs refer to the use of electromagnetic spectrum and network capacity. Whenever this term is used, the type of cost is explicitly specified as in the above examples. Financial (monetary) costs are usually linked to particular resource costs. For example, in public communication networks, there is usually a financial cost associated with the use of network capacity (communication cost). However, the relationship between financial costs and resource costs is highly variable and is influenced by non-technical factors such as business policy. Therefore, since the focus of this research endeavour is primarily scientific and technological in nature, considerations related to financial costs are beyond the scope of this work.

In order to obtain sufficient coverage of the major research in this field, a high-level systems approach is used throughout this work to abstract the in-depth details of individual research efforts. This does not imply that these details are unimportant but rather allows for the comparative analysis of multiple approaches in which different levels of detail are available. Due to the complexity of the technology involved and the focus of this minor dissertation, the physical implementation and testing of prototype mobile cloud computing systems is beyond the scope of this research. The research methodology used to identify and incorporate relevant sources of information in this work is discussed in **Section 1.6**.

It is envisioned that the outcomes and conclusions of this work will serve as a useful starting point for further research in the field of mobile cloud computing.

1.5 Dissertation Structure and Overview

Chapter 2 of this document explains the motivation for enhanced mobile computing using cloud resources. This chapter serves to contextualize the overall research endeavour through an investigation into the current situation and prevailing trends in the field of mobile computing.

Market research data shows that the number of mobile devices worldwide has increased rapidly over the past decade [6]. These devices can provide useful computational functionality through the use of mobile apps. Recently, both the range of available apps and the number of app downloads have exhibited rapid growth [7][8][12][13]. This combination of the increase in mobile devices and the rapid rise in mobile apps indicates a strong trend towards the use of mobile computing.

However, mobile devices will always be resource-constrained relative to non-mobile systems because of the fundamental constraints of mobility [14]. The device's small physical dimensions and finite energy storage capacity limit the choice of computational hardware and thus the computational capacity of the device. Wireless communication networks generally exhibit higher latency, lower bandwidth and less reliability than fixed equivalents. User

Interface (UI) capabilities are sometimes a limiting factor on small mobile devices.

It has been proposed that by connecting mobile devices to the cloud, the computational capacity of the mobile devices can be increased. Although the concept of offloading computation from resource-constrained devices is not new, three prerequisites for the practical realization of this idea have only recently been satisfied. These are the pervasive use of mobile computing, the broad availability of cloud resources and the widespread coverage of mobile communication networks. Therefore, it is now possible to achieve some degree of enhanced mobile computing through the use of cloud resources.

Chapter 3 presents a theoretical analysis framework for mobile cloud computing. Various applications of mobile cloud computing have been presented in recent literature. However the authors have generally used dissimilar criteria to analyse these proposals. The theoretical analysis framework defined in this chapter is a structured consolidation of the salient considerations identified in recent scientific literature.

By using a high-level qualitative approach, this framework includes a broad spectrum of analysis criteria. Three important requirements that guided the design of this framework are that it should be comprehensive in terms of its coverage, flexible in its application and enduring in its relevance.

The framework is composed of the following seven major aspects, each of which consists of several major considerations and key objectives:

- **Computational requirements** consist of the quantity, complexity and type of computational operations performed as well as the volume of data which is processed or stored.
- **Communication requirements** include the quantity of data to be transferred over the network as well as performance metrics such as communication bandwidth and latency.
- **Mobile network impact** analyses the impact of the system on a mo-

bile network and vice-versa. It compares communication requirements to mobile network capabilities.

- **Energy considerations** examine the energy costs and benefits for both the mobile device and the overall system. The system is assessed in terms of energy efficiency.
- **Information security** deals with issues of privacy and security arising from the transfer of data between the mobile device and the cloud over non-private networks.
- **System availability** investigates the risk of system failure and how the consequences of unavailability impact on the usefulness of the system.
- **Application usability** considers any limitations to applications due to the inherently constrained UI capabilities of mobile devices.

Through the process of grouping these considerations into the major aspects, various interdependencies have been identified. These indicate that certain aspects have an effect on one another. This validates the requirement for the use of a comprehensive analysis framework.

This framework can be used in the analysis of cloud-enhanced mobile applications. It facilitates comparisons between different systems based on particular aspects. Certain considerations from this framework can be used in the analysis of implementation approaches for mobile cloud computing systems.

Chapter 4 demonstrates the use of this framework in analysing various mobile application domains. A mobile application domain represents a specific purpose for which mobile devices can be used as computing tools. Each domain may consist of any number of relevant mobile apps. This grouping allows the domain-specific conclusions and recommendations to be generalized to multiple apps. These application domains are sourced from recent scientific literature as well as existing commercial endeavours. They have been selected based on the degree of benefit they can derive from the use of mobile cloud computing.

Some of these application domains have been discussed in recent scientific literature but have been analysed using only a limited subset of criteria. The use of the analysis framework provides a comprehensive qualitative analysis, supported by literature and examples, showing the impact of mobile cloud computing on each aspect of a domain. This identifies areas in which the domain would benefit from this technology as well as potential issues which must be addressed. It also allows for comparisons between different application domains based on specific aspects of the framework.

The following mobile application domains have been analysed:

- Mobile scientific computing and simulation
- Mobile electronic healthcare services (mHealth)
- Mobile tools for education and training (mLearning)
- Advanced mobile Human-Computer Interaction (HCI)
- Entertainment and multimedia services
- Mobile gaming applications

The analyses of these application domains using the analysis framework have resulted in various conclusions and recommendations specific to each domain as described in the relevant sections. They have also allowed for qualitative comparisons between different domains. The information obtained from these analyses can be fed back into the design process in order to enhance future applications using mobile cloud computing.

Chapter 5 presents an analysis of the various approaches to the implementation of mobile cloud computing systems.

Preceding the core analysis in this chapter are preliminary discussions of two relevant concepts. The first discusses the sub-paradigm of mobile grid computing. Ultimately, mobile grid computing is beyond the scope of this research due to the fundamental differences in system architecture between this concept and mobile cloud computing. The second preliminary discussion focuses on partitioning schemes for dividing a computational load between the mobile device and the cloud. Static partitioning is highly deterministic and simple to implement but is inflexible. Dynamic partitioning is more complex to implement but allows the system to adapt and compensate for device heterogeneity and mobile network fluctuations. As an open research area, new developments in partitioning will have an impact on mobile cloud computing.

The implementation approaches in this chapter have been selected from recent scientific literature as well as commercial endeavours. Although not an exhaustive list, this chapter includes a broad variety of approaches ranging from theoretical suggestions to commercial products. A high-level analysis is presented which draws on aspects of the theoretical analysis framework where relevant. Based on their architectural characteristics, the implementation approaches are grouped into three broad categories so as to obtain insight and conclusions which are applicable to a number of different systems.

- **Application-specific cloud services** use the cloud to provide predefined functionality for use by a specific application. Although this entry-level approach is the simplest to implement, it only enhances a single application and so has limited benefit to the mobile device as a whole.

- **Multi-application cloud services** provide limited cloud-based functionality to all applications on the mobile device. This type of mid-range approach is more complicated to design and implement than the application-specific approach. However, it provides benefit to the mobile device as a whole by enhancing a specific area of functionality across multiple applications.

- **Multifunctional cloud resources** currently represent the most advanced type of approach to mobile cloud computing. They provide augmented computational capacity to all mobile applications thus enhancing all aspects of the mobile device. Although they are the most complex to implement, approaches of this type provide an optimal solution for mobile cloud computing.

All three categories of implementation approaches have various advantages and disadvantages. In the short-term, it is likely that all three types of approaches will continue to co-exist and will be successfully used in different application domains. As the use of mobile cloud computing increases, it is likely that multifunctional cloud resources will become significantly more prominent in order to enhance system efficiency.

Chapter 6 presents the overall conclusions and summarizes this research. Predicted future trends in mobile cloud computing and recommendations for future work are also discussed in this chapter.

In terms of the future of mobile cloud computing, two major predicted trends have been identified. The first is that Mobile Network Operators (MNOs) will become increasingly involved in this technology since they can provide various technological benefits such as integrated services. The second is that the use of mobile cloud computing will continue to increase, leading to higher communication requirements and ultimately causing a shift towards the use of multifunctional cloud resources for reasons of efficiency.

Recommendations for future work include further experimentation and testing in the area of multifunctional cloud resources as well as in the initial association phase between device and cloud. The concept of cellular mobile cloud computing represents an important research opportunity. Investigation could be conducted into the use of the analysis framework in the analysis of mobile devices with minimal UI capabilities. It would also be informative to investigate the concept of mobile cloud computing from the cloud perspective. The final recommendation introduces the possibility of creating a mutually beneficial distributed computing relationship between mobile devices and the cloud by combining the vast computational resources of the cloud with the global decentralized network of mobile devices.

1.6 Research Methodology

This work primarily involves the consolidation, review and analysis of existing research and developments in the field of mobile cloud computing. This is similar to the concept of a meta-analysis as described by Rosenthal and DiMatteo [15]. However, in this work, the purpose of the analysis is to identify the salient considerations in the field and thus is not focused on a particular variable or set of variables.

This section presents the research methodology used to identify relevant sources of information and the criteria used in incorporating these into this work. The primary sources of information on which this work is based include both peer-reviewed research publications as well as active commercial endeavours. Due to the differences between these types of sources, the methodology used to identify relevant information and the criteria with which it is evaluated were adapted accordingly.

Research Publications: Peer-reviewed research publications such as journal articles and conference proceedings are usually organized and published in a relatively structured system. This facilitates textual searches based on keywords specific to the subtopic of interest. These searches were mainly directed towards online publication databases which broadly encompass this field including the *IEEE Xplore*[1] and *ACM Digital Library*[2] databases. Apart from textual searches, bibliographic links were also used as an important method in the identification of relevant research publications. Since mobile cloud computing is a relatively new field of technology, the majority of core sources were published within the last four years. Therefore, the date of publication was not used as a evaluation criterion. The primary criteria used in incorporating research publications in this work were the relevance to a particular subtopic and the level of detail available in the publication.

Metrics such as a publication's citation count and number of downloads were

[1] http://http://ieeexplore.ieee.org/
[2] http://http://dl.acm.org/

used as secondary criteria due to the relative recency of this field. Various research publications ranging from theoretical proposals to physical implementations have been used in this work.

Commercial Endeavours: The research publication sources are complemented by information about existing commercial systems which use mobile cloud computing. Since many of these systems are targeted at end-users, a significant amount of information is available via online sources. Although information about commercial endeavours is not available from structured databases, this information is often made available directly by the service provider. In comparison to research publications, there is less technical information available regarding commercial systems. The primary criteria for the incorporation of commercial endeavours in this work include the relevance to a particular subtopic and the level of detail available. A third criterion is that a fully functional production-quality system must be available. This effectively replaces the peer-review process and ensures that in all cases, the technological feasibility of these commercial systems has already been demonstrated.

Using these two types of sources, specific approaches and criteria were used to identify information relevant to the individual chapters of this work.

Motivation for Enhanced Mobile Computing: In order to achieve the first objective and to contextualize this work, quantitative information was obtained from research publications as well as recent business intelligence reports. The reputation of the publisher was an important criterion in the inclusion of business intelligence reports. Although these reports sometimes include speculative predictions, they provide an indication of overall trends within this field of technology as explained in **Chapter 2**.

Analysis Framework for Mobile Cloud Computing: Various considerations related to mobile cloud computing are presented and discussed in recent research publications. In some cases, the terminology differs between sources and was adapted accordingly. The major considerations which are common to multiple sources were consolidated and grouped into the aspects which form the theoretical analysis framework defined in **Chapter 3**.

Enhanced Mobile Application Domains: The objective of **Chapter 4** is to demonstrate the process of using the framework in the analysis of specific application domains. The application domains analysed in this section were selected primarily based on two criteria: Firstly, the degree of benefit the application domain can derive from the use of mobile cloud computing and secondly, the potential impact of the application domain (including its estimated user-base). This chapter is not intended to provide an exhaustive list of application domains but rather an illustrative sample in order to demonstrate the process of using the analysis framework. For each application domain, multiple sources were used to obtain information about each aspect.

Implementation Approaches for Mobile Cloud Computing: Various approaches to the implementation of mobile cloud computing systems are categorized and analysed in **Chapter 5**. The categorization was developed based on information from research publications as well as commercial endeavours. An important consideration for the inclusion of sources in this chapter was the level of technical detail available in the description of the system architecture. Since the systems in each category share the same architectural characteristics, advantages and disadvantages identified from a particular source can be applied to other systems in that category.

Overall, this research methodology facilitates the primary aim of this work which is to provide a consolidated review and structured critical analysis of current research and developments in the field of mobile cloud computing.

1.7 Conclusions

Through the completion of the objectives specified in **Section 1.4**, this research has resulted in the following outcomes and conclusions:

It has been shown that current circumstances present a very strong motivation for the use of enhanced mobile computing. The trend towards mobile computing has resulted in new mobile applications which place increasing computational demands on mobile devices. Due to the fundamental constraints of mobility, this demand cannot be met solely by hardware advancement. Through recent rapid developments, cloud resources have become widely available and are now well poised to meet these requirements by augmenting the computational capacity of mobile devices.

When analysing various systems based on the concept of mobile cloud computing, the interdependencies between different considerations make it critical to perform a comprehensive analysis. The theoretical analysis framework defined in this work achieves this by providing a structured consolidation of the major considerations relevant to mobile cloud computing which have been identified in recent scientific literature.

Various mobile application domains have been proposed as candidates to benefit from mobile cloud computing. By using the analysis framework, it is possible to obtain a comprehensive review of each domain as well as informative comparisons between different domains. Although this analysis has identified issues that need to be addressed in specific application domains, the overall conclusion is that mobile computing in these domains can be significantly enhanced through the use of cloud resources.

In terms of the implementation approaches to mobile cloud computing, it has been shown that single-application cloud services are the simplest to implement but only provide limited benefits. Multi-application cloud services are more beneficial but also more complex. Multifunctional cloud resources are the most complex but represent the optimal overall approach to mobile cloud computing. It is likely that future increases in the use of mobile cloud computing will ultimately cause a shift towards the use of multifunctional cloud resources in order to increase system efficiency.

Future trends in this area include increased involvement by MNOs and further increases in the use of mobile cloud computing. Various opportunities for future research exist within this field and these will be important in realizing the full potential of mobile cloud computing.

Chapter 2

Motivation for Enhanced Mobile Computing

2.1 Introduction

Building on the definitions in the previous chapter, mobile computing can broadly be defined as the use of mobile devices for computational purposes. In the context of this work the definition is restricted to include only the use of Mobile Computation & Communication Devices (MCCDs) as defined in **Section 1.3.1**. By this definition, mobile computing generally involves the use of devices such as high-end cellular phones, smartphones and tablet PCs.

A distinction can be made between mobile computing and the use of portable computing systems such as laptops or netbooks. Although portable systems can be used at different geographical locations, it is generally accepted that they will be used while stationary. Mobile devices are portable but are also capable of fully-functional operation whilst moving. Although the focus of this work is on mobile devices, some of the concepts presented are also relevant to portable computing devices.

Mobile computing is also related to the concept of ubiquitous computing.

Based on scientific literature, the concept of mobile computing precedes that of ubiquitous computing by a number of years. Mobile computing focuses specifically on the user's personal mobile devices. As defined in a seminal work by Mark Weiser, ubiquitous computing involves making many computers available throughout the physical environment, while making them effectively invisible to the user [16]. Although they share a common objective, mobile computing and ubiquitous computing exhibit fundamental differences in terms of overall system architecture. This work focuses on mobile computing using MCCDs but some of the principles discussed in this work are also applicable to ubiquitous computing.

The purpose of this chapter is to contextualize the research by providing background information about aspects relevant to mobile cloud computing. In their 1994 paper, Forman and Zahorjan anticipated that mobile computing *"will very likely revolutionize the way we use computers"* [17]. Today, this potential is being realized through the pervasive nature of MCCDs and the rapid increase in mobile software applications. **Section 2.2** explores this prevailing trend towards mobile computing.

There are a number of challenges which are inherent to mobile computing. These are mainly caused by the limitations of mobile devices. Although the hardware capability of these devices is constantly improving, the rate of improvement is limited by the constraints of mobility and other considerations as explained in **Section 2.3**.

As introduced in the previous chapter, the cloud computing paradigm is undergoing a period of rapid technological advancement and diversification. **Section 2.4** provides further background information about this paradigm by exploring the current capabilities of cloud resources.

Ultimately, the objective of mobile cloud computing is to utilize these cloud resources to augment the computational capacity of the mobile device in order to enhance mobile computing. Although this type of proposal is not new, the conditions necessary for its successful implementation have only recently been satisfied as explained in **Section 2.5**.

2.2 The Trend Towards Mobile Computing

The International Telecommunication Union (ITU) World Telecommunication/ICT Indicators Database provides time-series data for various telecommunication and Information & Communication Technology (ICT) developments from around the world [18]. This data is sufficiently reliable because it has been obtained from official contacts in each country such as the ministry in charge of telecommunication [18]. The Global ICT Developments for the years 2000 to 2011 from this database are shown in **Figure 2.1** [6].

As shown in the lower right of the figure, there were 17.0 active mobile broadband subscriptions per 100 inhabitants worldwide in 2011 compared to 8.5 fixed (wired) broadband subscriptions in same year. The growth rate of mobile broadband subscriptions has also exceeded that of fixed broadband subscriptions for the indicated period.

The uppermost line in the figure shows the significant growth in mobile cellular telephony which had 86.7 subscriptions per 100 inhabitants in 2011. Since this figure shows active subscriptions, it follows that all these users must have mobile devices. Based on the shape of this plot, it appears likely that the number of mobile devices will continue to increase in the near future. As technology advances, one of the key aims will be to extend the reach of mobile broadband connectivity to these mobile users. Since 2005, the number of fixed telephone subscriptions has steadily decreased. This data shows a global trend towards the use of mobile devices for communication and broadband internet connectivity.

Mobile computing involves the use of mobile devices for computational purposes. As a concept, mobile computing is not a new idea. In 1996, Satyanarayanan published a now widely-cited paper exploring the *"Fundamental Challenges in Mobile Computing"* [14]. Although technologies have changed significantly, most of the fundamental constraints identified in that paper are still relevant today as discussed in **Section 2.3**.

In the past, mobile computing functionality was provided by Personal Digital

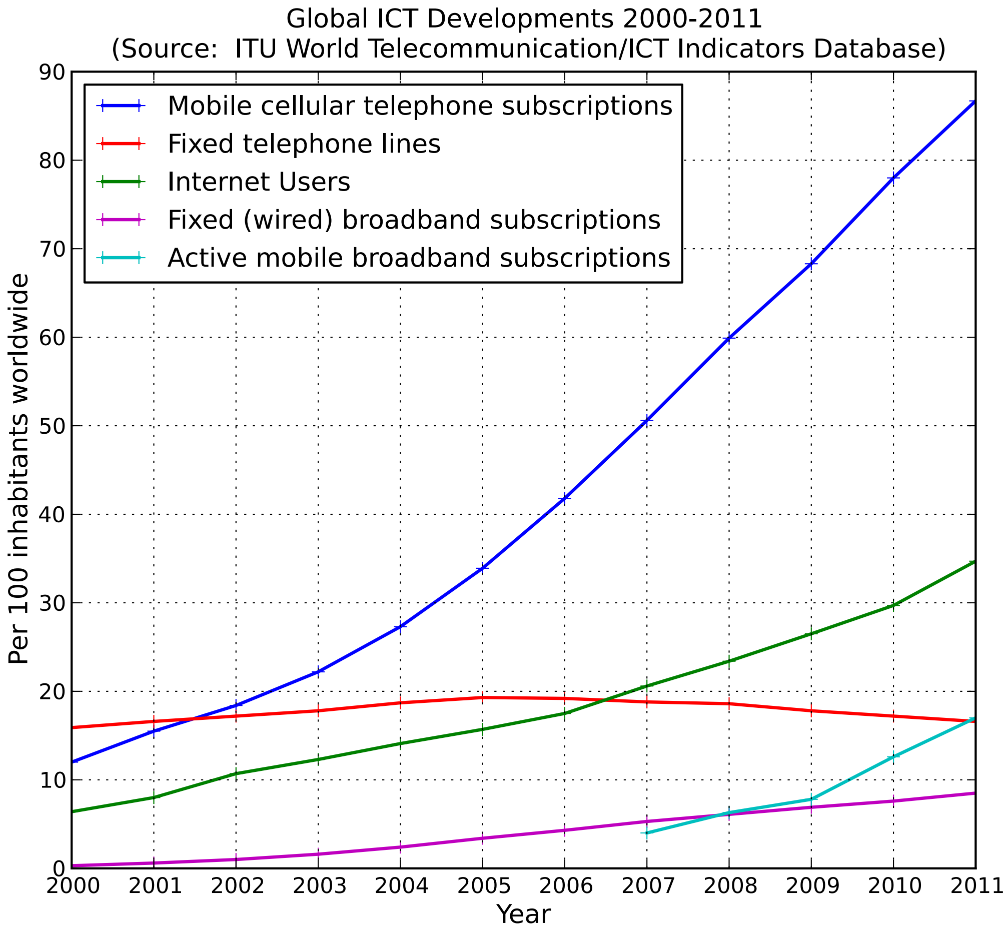

Figure 2.1: Global ICT developments for the years 2000-2011 [6] obtained from the International Telecommunication Union (ITU) World Telecommunication/ICT Indicators Database [18]. This graph shows the worldwide number of subscriptions/users of each technology normalized as a percentage of the global population. This data has been obtained by the ITU from official contacts in each country. Data for the year 2011 has been estimated by the ITU.

Assistants (PDAs) but in recent years, the PDA has been almost completely superseded by a range of MCCDs. MCCDs such as smartphones and tablet PCs arguably evolved from a combination of mobile cellular phones and PDAs in terms of their form-factors and functionality. These devices now provide significant computational capacity which can be used to run mobile software applications (commonly referred to as *'mobile apps'*).

Recent mobile app usage data show that there is an extremely high growth rate in this sector. Gartner anticipates 17.7 billion mobile apps will be downloaded in 2011 [8]. This is more than double the number of downloads in 2010. The Gartner report also forecasts that revenue derived from these applications will surpass USD 15.1 billion in 2011 which is an increase of 190% over the 2010 revenue figure [8]. It is predicted that between 2010 and 2014, this sector will grow to more ten times its current size and generate revenues of approximately USD 58 billion [8]. In terms of the number of new mobile apps developed, data from IDC shows that over 300,000 different apps have been developed in the three year period 2008-2010 [7]. Similar growth predictions have been made by In-Stat [12] and Juniper Research [13].

This data shows that the use of mobile apps is increasing rapidly. Since mobile apps utilize the computational capacity of the mobile device to provide useful functionality to the user, their use is a form of mobile computing. Therefore, this shows that mobile devices are increasingly being used as computing tools. The combination of the increasing use of mobile apps and the increasing number of mobile devices illustrates a clear trend towards mobile computing.

MCCDs provide the capability and capacity required for mobile computing including computation, storage, communication and user interface functionality. Furthermore, these devices generally have a lower financial cost than traditional PCs and have become readily available as mainstream products.

Although mobile devices have now become pervasive, there are still a number of challenges to mobile computing. These are mostly caused by the inherent limitations of mobile device hardware as explained in the next section.

2.3 Limitations of Mobile Devices

Satyanarayanan has proposed a number of constraints which are intrinsic to mobility [14]. These constraints are:

- Mobile systems are resource-constrained relative to non-mobile systems using similar technology [14].
- Mobile systems rely on an energy source with a finite energy storage capacity [14].
- Mobile network connectivity is highly variable in terms of performance and reliability [14].

In today's context, MCCDs are still more limited than non-mobile systems such as PCs in terms of data processing and storage capacity as well as network connectivity.

At present, the computational capacity of an MCCD is almost exclusively determined by the computational hardware of the device and therefore it is possible to increase this capacity by using more advanced hardware. However, the choice of hardware which can be used in mobile devices is limited by the fundamental constraints of mobility. Based on the definition of an MCCD from the previous chapter, these hardware limitations arise from the finite energy storage capacity and the constrained physical size of the device. In addition, financial costs also restrict the choice of mobile device hardware.

A general characteristic of computational hardware is that higher capacity hardware has higher power requirements. For example, more power is required to achieve faster processor clock speeds or to support multiple processing units. Due to the finite energy storage capacity of the mobile device, any increase in power consumption will reduce the device's mobile runtime which is commonly, but incorrectly, referred to as the device's *'battery life'*. A survey conducted in September 2011 showed that nearly 60% of users rank battery life (mobile runtime) as an essential consideration when buying a new mobile device [19]. This online survey of over 30 000 mobile phone and tablet users found that, overall, battery life (mobile runtime) is more important than the screen size, ease of use, appearance or price of the mobile device [19]. Current battery technology is limited in terms of energy density and it will be challenging to increase this energy density in the near future using current technologies. A report published by the National Research Council (of the United States of America) states that "using new materials and chemistries, batteries are approaching explosives in terms of energy density" [20].

Without any significant improvements in battery energy density, increasing the energy storage capacity of a battery will result in a proportional increase of its physical size. In the context of mobile devices, this is usually infeasible due to the constrained physical dimensions of the device. As a result of its increased power requirements, higher capacity computational hardware will generate more heat and so require greater heat dissipation within the device. The mobile device's physical size limits the amount of heat which it can dissipate. Apart from computational hardware, MCCDs also include other hardware subsystems such as the various communication subsystems, the user interface and any auxiliary subsystems such as sensors. Therefore, the physical size of the computational hardware itself is also a limiting factor since it must share space with these other hardware subsystems within the size-constrained mobile device. For example, data storage density is an important consideration in non-volatile data storage technologies such as solid-state memory. Without increasing storage density, an increase in storage capacity will result in an increase in physical size.

The choice of hardware which can be used in a mobile device is also limited by financial considerations. Highly specialized computational hardware usually carries a significantly greater financial cost. Since MCCDs are usually mainstream devices, the use of such advanced hardware would not be financially viable. To some degree, the high production volumes of mobile devices serve to reduce financial costs through economies of scale.

In terms of communication, the mobile wireless communication technologies used by MCCDs are more limited than fixed equivalents. Mobile networks are usually characterized by lower bandwidths and higher latencies than non-mobile equivalents [17]. The maximum power which the mobile device may radiate for wireless transmissions is often limited by legislation. However, from a technological perspective, the amount of power available for wireless communication is fundamentally limited by the finite energy storage capacity of the mobile device.

In mobile computing, physical size constraints generally limit the capabilities of the device's User Interface (UI). This constrained UI in turn limits the functionality that can be provided by the mobile device. For example, previous generations of mobile phones only featured small monochromatic screens and basic numeric keypads. Although these devices were designed primarily for mobile communication rather than mobile computing, any use of such devices for mobile computing would have been severely limited by the UI constraints. Although UI limitations are an inherent challenge in any mobile computing system, modern MCCDs have undergone significant improvements in this area. The UI of a modern MCCD generally consists of a large full-colour touchscreen display which provides a dynamic combination of input and output capabilities. This maximizes spacial efficiency because the entire display area can be used for input or output or a combination thereof as required by the application. For example, the on-screen keyboard of a tablet PC can fill the screen when a user is entering text and be completely hidden when graphical output is displayed. In some applications, insightful UI designs such as touch-gesture recognition can utilize the full screen area for input and output simultaneously. UI design for mobile devices is an active area of research and it is anticipated that even more efficient UI designs will become available as the trend towards mobile computing continues.

As new applications are developed, it is increasingly likely that their requirements will exceed the computational capacity of mobile devices. In order to continue to support the trend towards mobile computing, it is necessary to augment this capacity. Although advances in device hardware will continue to contribute to this objective, the rate of improvement is limited by the fundamental constraints for mobility and so will not be sufficient to support the current growth in application requirements. Therefore, in addition to ongoing hardware development, there is also scope for further enhancement of mobile computing using alternative approaches such as mobile cloud computing.

2.4 Characteristics of Cloud Resources

Cloud computing has recently undergone a period of rapid technological development. Although definitions of this emerging technology still vary widely, many are beginning to converge on the views presented in the 2009 work by Armbrust et al. titled *"Above the Clouds: A Berkeley View of Cloud Computing"* [10]. The IEEE has begun the process of formulating standards for cloud portability and interoperability [21].

Two fundamental characteristics of cloud computing which are universally accepted in recent literature are that:

- The cloud paradigm provides very high-capacity (seemingly unlimited) computational resources [9][10].
- These resources are available on-demand with a very high degree of elasticity [9][10].

The realization of each of these characteristics is explained in **Section 2.4.1** and **Section 2.4.2** respectively. Further discussion regarding the scope of cloud resources in the context of this work is provided in **Section 2.4.3**.

2.4.1 High-Capacity Computational Resources

In order to achieve the first of these characteristics, cloud providers usually utilize large-scale data centre infrastructure. This high computational capacity is feasible because of economies of scale [10]. Hamilton has shown that the financial costs of hardware, network bandwidth and power are reduced by a significant factor when operating a large-scale data centre [22]. Fixed costs such as administration can also be distributed over a larger number of servers [22]. **Figure 2.2** shows a graphical representation of the benefits of large-scale data centres. This figure is based on data from Hamilton for the year 2006 [22]. Although the capacity of computational infrastructure has increased significantly in the last five years, this figure is still relevant because it shows the relationship between data centres of different sizes. As shown in **Figure 2.2**, a large-scale data centre can obtain 7.3 times more network bandwidth and 5.6 times more storage capacity than a mid-sized data centre for the same financial cost [22]. Administration costs per server are at least 7.1 times lower in the large-scale data centre [22].

From a technological perspective the computational architectures of cloud computing systems are designed to achieve high levels of efficiency. Various internet and software companies such as Google [23], Amazon [24] and Microsoft [25] now provide cloud services. Even a small percentage of this infrastructure would be considered a virtually infinite resource in comparison to the computational capacity of a mobile device.

2.4.2 High Degree of Elasticity

In the context of cloud computing, the term *'elasticity'* is frequently used to refer to the scalability of the computational systems, the flexibility with which these systems can be scaled up or down and the time required to achieve this. Therefore, in a highly elastic system, computational capacity can be scaled very rapidly (in the order of seconds). Elasticity is largely realized through extensive use of virtualization technology. Virtualization

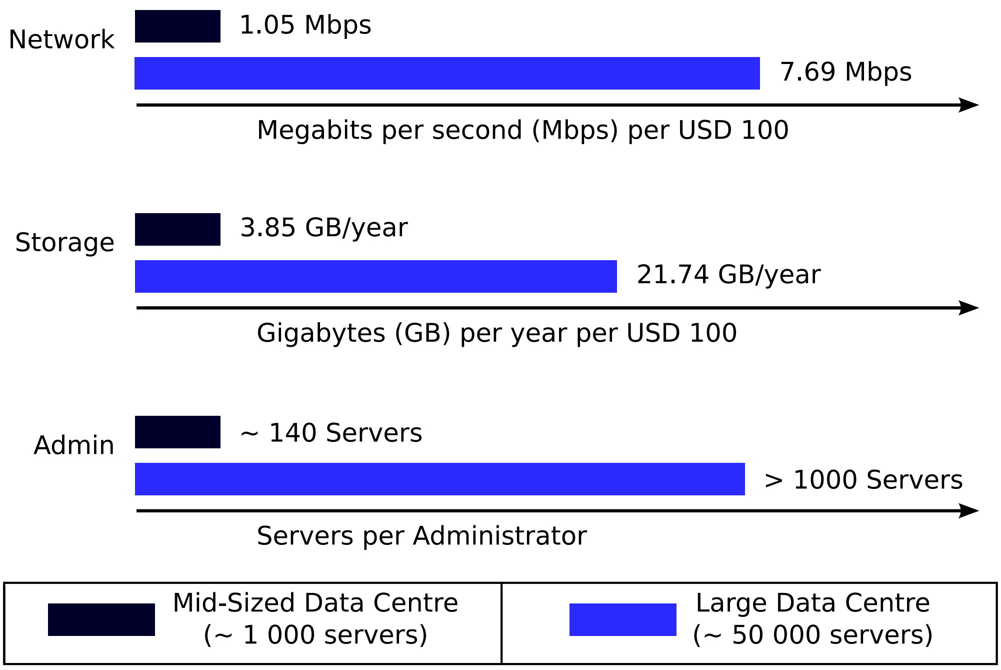

Figure 2.2: A graphical representation of the comparison between a mid-sized data centre and a large-scale data centre based on data obtained by Hamilton for the year 2006 [22]. This data shows that a large-scale data centre can obtain 7.3 times more network bandwidth and 5.6 times more storage capacity than a mid-sized data centre for the same financial cost [22]. As explained by Armbrust et al., the high computational capacity of cloud computing infrastructure is feasible due to these economies of scale [10].

is the process of decoupling computational resource units from the underlying hardware systems. This allows the available hardware infrastructure to be dynamically partitioned and allocated as required. Since this process is controlled in software, the time required for provisioning or removal of virtualized computational resources is in the order of seconds. By using sufficiently fine-grained control for the provision of resources, the situation of over-provisioning (and therefore under-utilization) can be avoided. This serves to increase the overall energy efficiency of the system since the available hardware infrastructure can be utilized at its optimal capacity. Ongoing research efforts such as *"Energy Aware Consolidation for Cloud Computing"* by Srikantaiah et al. address the problem of consolidating virtualized resources on shared hardware infrastructure [26]. The proposed solution defines the optimal utilization of computational infrastructure as the point at which the energy usage per computational operation is minimized [26].

2.4.3 Further Cloud Definitions

As explained in **Section 1.3**, for the purpose of this work it is sufficient to model the cloud as a remote resource which provides a dynamic combination of data processing, data storage and network connectivity. Resources complying with this definition can be provided by any of the current deployment models for cloud infrastructure (public, private, community or hybrid clouds). In order to accommodate the variety of application domains and implementation approaches analysed in subsequent chapters, this work does not focus on any specific cloud platform or technology. However, as explained in the relevant sections, the choice of cloud technology is important because it can have a significant impact on certain areas of an application. Recently, it has been proposed that a distinction should be made between a *'Business Cloud'* and a *'Personal Cloud'* as discussed by Tian et al. [27]. However, at this stage the difference between these is based almost exclusively on the types of applications supported by each cloud and does not relate to the technological aspects of the system. Therefore this distinction is not critical to this work and so the term *'cloud'* is used to refer to both personal and business clouds.

2.5 Advent of Mobile Cloud Computing

It has been proposed by various authors that the computational capacity of a mobile device can be enhanced through the use of cloud resources [28][29][30][1]. Fundamentally, mobile cloud computing involves a heterogeneous distributed computing relationship between a resource-constrained mobile device and a remote high-capacity cloud resource via a network such as the internet.

The aim of mobile cloud computing is to maximize both the performance and efficiency of the overall system. In this context, *'performance'* refers to the manner in which the system achieves its objectives. This concept is dependent on the application domain and could include metrics such as execution time and output quality (computational performance) as well as data transfer rate (communication performance). In this context, *'efficiency'* refers to the ratio between the useful output of the system and the resources consumed in achieving this output. This concept is also domain-specific and could include factors such as the amount of energy required (energy efficiency) or the amount of network data transfer needed (communication efficiency) for the system to complete a particular task. The specific performance and efficiency criteria for each application domain are discussed in the analyses presented in **Chapter 4**.

The underlying concept of offloading computation from resource-constrained devices is not a new idea. It is based on the principles of distributed systems and the client server model which have existed since the start of networked computing. However, the following three conditions, which are critical for the practical realization of this idea, have only recently been satisfied:

- The pervasive use of mobile computing systems.
- The broad availability of cloud-based computational resources.
- The widespread coverage of mobile communication networks.

Firstly, as shown in **Figure 2.1** and explained in **Section 2.2**, increasing numbers of mobile devices as well as mobile apps indicate a trend towards mobile computing. This creates a demand for computational capacity on mobile devices which will be challenging to meet by relying only on advances in mobile hardware.

Secondly, the cloud resources which provide the required computational capacity have become widely available. One of the descriptions of cloud computing given in the Berkeley paper is, *"a long-held dream of computing as a utility, which has recently emerged as a commercial reality"* [10].

Thirdly, the mobile communication networks which facilitate communication between MCCDs and the cloud have become available worldwide. These networks include Wide Area Networks (WANs), Metropolitan Area Networks (MANs), Wireless Local Area Networks (WLANs) and even Personal Area Networks (PANs). Although the performance of mobile communication networks will continue to improve, Cuervo et al. have demonstrated that mobile cloud computing systems can already be supported using current mobile communication technologies [31].

Therefore, it is now feasible to propose that, to some degree, the limited computational capacity of mobile devices can be augmented through the use of cloud resources.

2.6 Conclusion

The purpose of this chapter is to contextualize the overall research endeavour and to present the motivation for enhanced mobile computing using cloud resources.

It has been shown that there is currently a trend towards the use of mobile computing. The ITU World Telecommunication/ICT Indicators Database shows that the number of mobile devices worldwide has increased rapidly over the past ten years [18]. Given this trend, it is likely that the number of mobile devices will continue to increase in the near future. This database also indicates that the number of mobile broadband connections has exceeded that of fixed broadband connections since 2008 [6]. Data from various business intelligence reports published by the IDC [7], Gartner [8], In-Stat [12] and Juniper Research [13] have shown rapid growth in the number of applications developed for mobile devices. This combination of increasing numbers of mobile devices coupled with the rapid rise in mobile applications indicates a trend towards the use of mobile computing.

Mobile devices are resource-constrained relative to non-mobile systems because of the fundamental constraints of mobility. At present, the computational capacity of an MCCD is determined by the computational hardware of the device. The choice of mobile device hardware is limited by the finite energy storage capacity and the constrained physical size of the device. Since these are mainstream devices, financial considerations also limit the choice of hardware for mobile devices. The wireless communication networks used by mobile devices are more limited than the fixed equivalents in terms of bandwidth, latency and reliability. Although the small physical size of the device still poses a challenge in terms of mobile UI capabilities, new research and development efforts are providing promising solutions to overcome this limitation.

Two fundamental characteristics of the emerging cloud computing paradigm are that it offers very high-capacity computational resources and that these resources can be provisioned on-demand in a highly elastic manner. Cloud providers can use large-scale data centres to provide these high-capacity computational resources. By leveraging economies of scale, the use of high-capacity hardware infrastructure can increase efficiency and minimize the financial cost of computation. The high degree of elasticity in the cloud is realized through the use of virtualization technology. Virtualized resources can be provisioned and removed in a matter of seconds. This decoupling of the computational resource units from the underlying hardware makes it possible to operate the hardware infrastructure at its optimal capacity, thereby maximizing the overall efficiency of the system.

Mobile cloud computing has been proposed as a means of enhancing mobile computing through the use of cloud resources. Using this paradigm, a resource-constrained mobile device establishes a distributed computation relationship with remote high-capacity cloud resource via a network such as the internet. The cloud resource is used to augment the computational capacity of the mobile device, increase its performance and improve the efficiency of the system. Three important conditions for the realization of this concept have recently been met. Firstly, the pervasive nature of mobile devices and the trend towards mobile computing has created a demand for mobile computational capacity which will be challenging to meet using only hardware advancements. Secondly, cloud resources, which have become broadly available, can provide the additional computational capacity required by these mobile devices. Thirdly, the widespread coverage of mobile communication networks provides the means by which mobile devices can interact with the cloud. Therefore, it is now possible to achieve some degree of enhanced mobile computing through the use of cloud resources.

Chapter 3

Analysis Framework for Mobile Cloud Computing

3.1 Introduction

Although it is still an emerging technology, mobile cloud computing has been discussed in a number of publications in recent scientific literature. Some of these publications focus on applications of this technology whilst others are concerned with the implementation of the underlying systems. These publications range from theoretical proposals through to prototype implementations supported by quantitative measurements and benchmarks.

However, when analysing various mobile cloud computing systems, the authors of these publications have used dissimilar subsets of criteria in their analyses. Although this selective analysis is often useful in emphasizing the primary focus of the work and communicating findings, it makes it difficult to obtain a holistic view of the situation. Comparisons between different proposals or research efforts are limited by the lack of common analysis criteria.

Simoens et al. have reviewed and analysed a number of mobile cloud computing systems specifically in the area of remote graphical rendering [11].

They have proposed that future research should focus on *"the design of an overall framework"* for mobile cloud computing [11].

The aim of this chapter is to identify the various considerations relevant to mobile cloud computing based on recent literature and to consolidate these into a structured theoretical analysis framework. In the context of this work, an *'analysis framework'* is defined as a theoretical construct consisting of a number of distinct aspects used for analysis purposes.

Three important requirements of such an analysis framework are that it should be comprehensive in terms of its coverage, flexible in its application and enduring in its relevance. Firstly, in order to be considered comprehensive, the analysis framework must address all the important considerations relevant to this technology. In cases where multiple considerations are closely related to each other, these can be grouped into major aspects. Secondly, the flexibility of the framework determines its usefulness as an analysis tool. The framework should therefore be applicable to different types of analyses within the field of mobile cloud computing. For example, it should be sufficiently flexible to be used across various application domains. Thirdly, given that technology will continue to develop, it is important that the framework will remain relevant and useful in the future. Whilst this is impossible to ensure, the endurance of the framework can be improved by focusing more on fundamental considerations which are largely independent of any particular technology.

In terms of level of detail, an analysis framework can vary from a high-level conceptual overview to a detailed low-level approach. Although both options meet the above requirements, the various metrics and benchmarks included in a low-level approach can be used for quantitative analysis whereas the high-level approach is inherently limited to a more qualitative conceptual analysis. However, the major disadvantage of a low-level approach is that it requires constant maintenance in order to remain relevant. As technology develops, the low-level metrics and benchmarks will inevitably need to be modified and updated to include new reference levels and measurement techniques.

Due to the focus of this minor dissertation and the fact that it is a single-pass design rather than an ongoing effort, the analysis framework defined in this chapter uses a relatively high-level approach. Since it does not depend on low-level quantitative metrics, it will theoretically remain relevant for longer than a low-level quantitative approach.

This analysis framework consists of seven major aspects, each representing a category or similar logical grouping of considerations. Although some research efforts may use more exotic nomenclature for their analysis criteria, the aspects in this framework are sufficiently broad to encompass the fundamental considerations in the field of mobile cloud computing. The seven major aspects which constitute this framework are:

- **Computational Requirements**
- **Communication Requirements**
- **Mobile Network Impact**
- **Energy Considerations**
- **Information Security**
- **System Availability**
- **Application Usability**

Through the process of grouping the considerations into the major aspects, various interdependencies between aspects have been identified. These indicate that certain aspects have an effect on one another. This validates the requirement for the use of a comprehensive analysis framework.

Detailed descriptions of the aspects of this framework are presented in the subsequent sections of this chapter. A summary of the interdependencies between aspects is given in **Section 3.9**. In **Chapter 4**, the aspects identified in this framework are used to analyse and compare various mobile application domains. In **Chapter 5**, some of these aspects are used in the analysis of various implementation approaches for mobile cloud computing systems.

3.2 Computational Requirements

As discussed in **Chapter 2**, there is currently a trend towards the use of mobile computing. The functionality of a mobile computing device is usually provided by mobile software applications. Since these applications perform some degree of computation on the mobile device, they can be said to have a definable set of computational requirements. This concept is analogous to the PC domain where software often has *'minimum system requirements'*.

At a fundamental level, a mobile computing device must be capable of processing and storing some quantity of data. As expected, data storage capacity is a measure of the quantity of data which can be stored persistently on the mobile device (number of bits). Processing performance is a measure of the rate at which the device can perform computational operations (number of operations per second). From a hardware perspective, processing performance depends on the architectures and clock speeds of the various processing units on the device as well as the quantity of Random Access Memory (RAM) available. Fundamentally, the overall computational capacity of the mobile device can be represented as a combination of the device's processing performance and data storage capacity.

As explained in **Chapter 2**, the computational capacity of a mobile device is limited by the fundamental constraints of mobility. These constraints limit the performance and output quality of mobile applications. In some cases, the minimum computational requirements of an application exceed the computational capacity of the mobile device making it impossible to provide this functionality without increasing the capacity of the mobile device.

The mobile cloud computing paradigm aims to remove this limitation by augmenting the computational capacity of mobile devices through the use of cloud resources. When required, the mobile device can establish a heterogeneous distributed computing relationship between itself and its associated cloud resource via a network such as the internet. This architecture is represented diagrammatically in **Figure 3.1**.

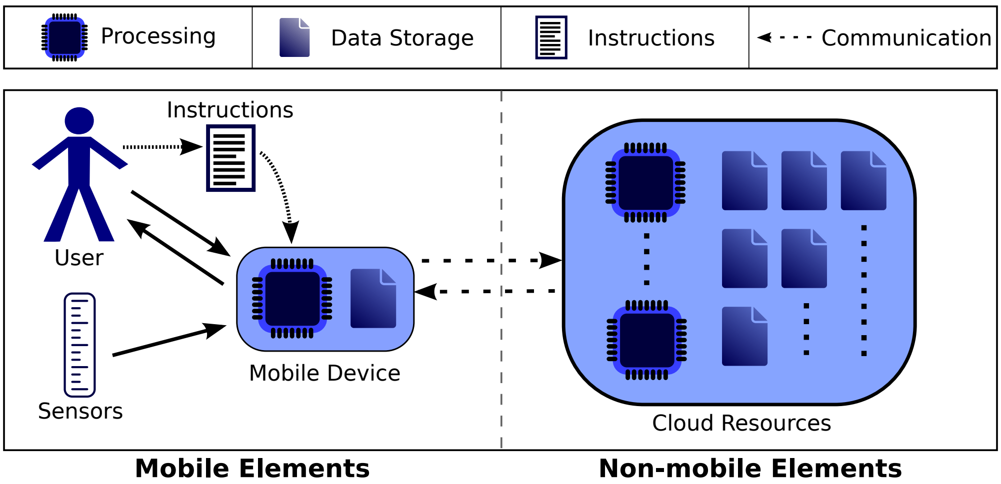

Figure 3.1: A high-level overview of the computational architecture of a mobile cloud computing system. The mobile device is the primary point of interaction with the system. The user interacts directly with the mobile device by entering instructions (commands) and data. The device also obtains input data from various peripheral sensors. The limited computational capacity of the mobile device (processing and data storage) is augmented through a distributed computing relationship with the remote high-capacity cloud resources via a mobile wireless communication network.

In order to assess the benefit provided by mobile cloud computing from a computational perspective, a naive approach would be to measure the increase in raw computational capacity. However, since the capacity of the cloud is usually at least an order of magnitude greater than that of the mobile device, this would simply be a measurement of the capacity provided by the cloud. A better approach is to compare the computational performance of specific applications to a baseline reference obtained without the use of the cloud resource. This comparison can utilize appropriate metrics such as the time taken to complete a specific operation (execution time) or the quality of the output obtained (sometimes referred to as the fidelity of the output).

Kun, Shumao and Hsiao-Hwa have demonstrated that the use of a cloud resource reduces the execution time of an application for translating sentences into another language [32]. Using their *"MAUI"* mobile cloud computing system, Cuervo et al. have analysed a facial recognition application, a graphically-intensive video game and a chess game by measuring the execution time of specific scenarios [31]. Their measurements have also shown that the execution time of specific applications can be significantly reduced using mobile cloud computing [31].

As part of their *"CloneCloud"* project [33], Chun et al. have performed comprehensive execution time measurements on three different types of applications: virus scanning, image search and behaviour profiling [2]. By varying the quantity of data to be processed in each application, they have concluded that larger workloads benefit more than smaller tasks from offloading computation to a cloud resource [2]. This shows that the quantity of computation is an important consideration.

The idea of variable application fidelity (quality) has been described by Narayanan, Flinn and Satyanarayanan as the basis of the *"Odyssey"* system [34]. This system determines the appropriate level of fidelity for an operation based on a number of input parameters including resource consumption [34]. These same authors have also demonstrated the operation of *"Odyssey"* as part of a larger system called *"Spectra"* using a speech recognition application [35]. In the latter scenario, the application's execution time was measured and the results show how this time varies depending on the fidelity of the output [35]. The output quality is usually linked to the quantity and complexity of computation performed by an application. Therefore, the quantity and complexity of the required computational operations are important considerations in this aspect.

Another consideration which falls within this aspect is the type of computation required. Since the mobile device and the cloud resource essentially form a heterogeneous parallel computing system, the degree of computational parallelism of tasks is important. Furthermore, through the use of virtualization technology, the cloud may appear to be a single high-capacity computational resource but in reality the underlying hardware is likely to be some type parallel system. Therefore, as applications begin to utilize more of the augmented capacity provided by mobile cloud computing, Amdahl's law must be taken into consideration [36]. Amdahl's law states that for a particular operation, the increase in computational performance provided by a parallel computing architecture is limited by the percentage of the operation which cannot be parallelized [36]. This principle is often used to predict the theoretical maximum speedup which can be achieved by using a parallel computing architecture in a specific application. In general, mobile applications will include some operations which cannot be parallelized and thus Amdahl's law will place an upper limit on the performance increases obtained using mobile cloud computing. Marinelli has taken this into consideration in his work on the *"Hyrax"* system [3].

The cloud resource can also be used to augment the data storage capacity of the mobile device. Commercial services such as the *"Amazon Simple Storage Service"* (*"Amazon S3"*), can be used to store information in the cloud on behalf of the mobile device [37]. This cloud storage capacity is usually significantly greater than that of the mobile device. As explained by Kumar and Lu, the ability to store data persistently in the cloud is a fundamental differentiating factor between mobile cloud computing and the use of remote computing from mobile devices [1]. This persistent data storage can result in a significant increase in system efficiency because the same data does not need to be re-uploaded for later sessions as is often the case in the remote computing paradigm [1].

An overview of the principle contributions of this aspect to the overall analysis framework is shown in **Table 3.1**.

Table 3.1: Summary of the Computational Requirements Aspect

Major Considerations: • The quantity, complexity and type of computational operations required to complete a task. • The quantity of data to be processed or persistently stored. • The increase in computational performance of specific tasks when using a cloud resource. • The quality (fidelity) of the output obtained using a cloud resource.
Primary Objective: • Maximize the computational performance of the mobile device.

3.3 Communication Requirements

This aspect analyses the communication requirements of mobile cloud computing and demonstrates how these requirements affect the overall performance of systems based on this concept.

In previous generations of mobile computing, the mobile device was a self-contained computing system and so network connectivity was only a peripheral concern. However, mobile network connectivity is a fundamental part of the mobile cloud computing paradigm. In order to use mobile cloud computing, devices require mobile communication capability over and above their mobile computing capability. The vast majority of MCCDs in use today have evolved from cellular telephones and so have maintained a strong focus on mobile communication supported by third and fourth generation (3G and 4G) cellular network technologies.

In a mobile cloud computing system there are various communication links between the different entities. At a fundamental level, the most important of these are represented by the data-flow diagram shown in **Figure 3.2**.

As shown in **Figure 3.2**, the mobile device communicates with the cloud resource and other network end-points via wireless communication links. On the fixed side, cloud resources and other network end-points communicate using high-capacity fixed communication infrastructure. In comparison to fixed networks, wireless networks are generally more limited in terms of bandwidth and latency [17]. The availability of electromagnetic spectrum has become a limiting factor in terms of the capacity of wireless communication networks whereas fixed communication systems are not subject to this constraint. Therefore, although the fixed links are considered, the primary focus of this aspect is on the wireless communication links shown in **Figure 3.2**.

In the context of mobile computing, the mobile device is the primary source of data within the system. As shown in **Figure 3.2**, data could originate from user input or from any of the sensor subsystems on the mobile device.

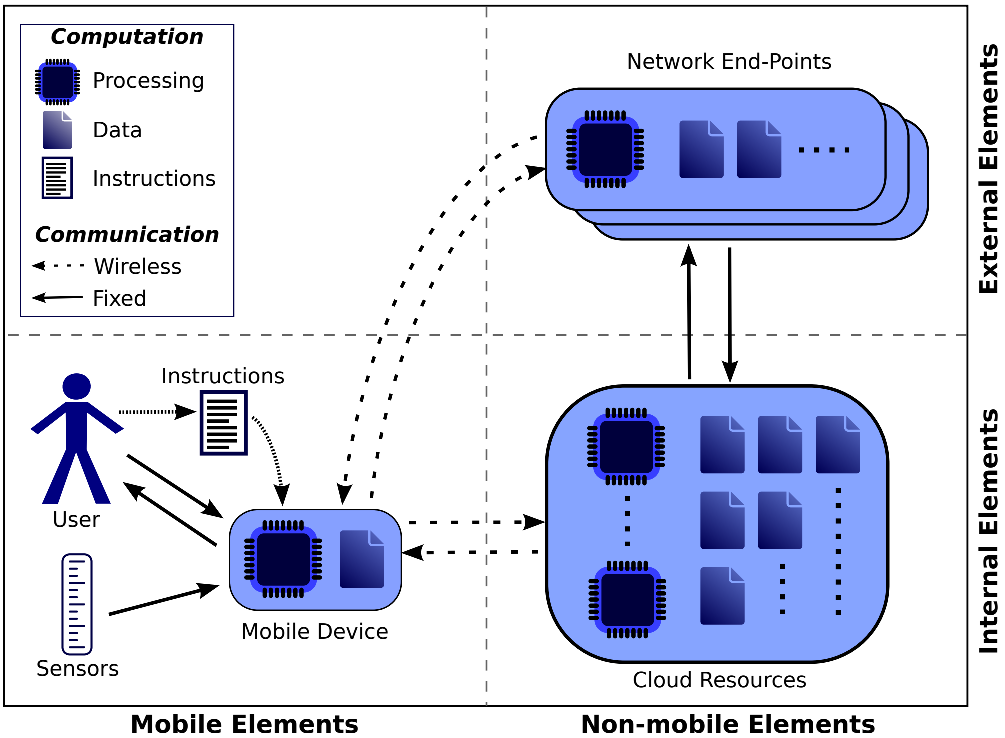

Figure 3.2: A high-level overview of the major functional elements and communication links in a mobile cloud computing system. The mobile device is the primary source and destination of data. To a lesser extent, the external network end-points are also sources and destinations of data. The cloud resources are classified as internal elements since they are completely controlled by the user. This distinguishes them from other fixed network end-points which do not afford the user full functional control.

In this context, the term *'sensors'* refers to all subsystems or peripherals of the mobile device which generate data. The range of such subsystems is expanding rapidly but some current examples include cameras, Global Positioning System (GPS) receivers, accelerometers and Near Field Communication (NFC) transceivers. A user can control the system by issuing commands (instructions) through the mobile device's UI. The mobile device is also the primary destination of data in the system since this is usually output to the user.

Although the cloud resource has significantly greater computational capacity than the mobile device, the latter is both the primary origin and destination of data. In order to utilize the computational capacity of the cloud resource, data and instructions must be transferred from the mobile device to the cloud via a wireless link. All increases in computational performance discussed in the previous section are dependent on the performance of this crucial communication link. The extent of this dependence is directly proportional to the quantity of data exchanged between the mobile device and the cloud. Therefore, there is an interdependency between the computational requirements and the communication requirements aspects of this framework.

In their work on parametric analysis for adaptive computation offloading, Wang and Li have defined an energy-based cost model to determine the optimum balance between local processing and computation offloading [38]. In this model, the data communication cost is directly proportional to the quantity of data transferred between the mobile device and the remote computational resource [38]. The objective function of the model aims to minimize the total energy cost which, from a communication perspective, requires the quantity of data transferred between the device and the cloud to be minimized [38].

In addition to the quantity of data transferred, the required rate of this transfer defines the bandwidth requirements for this communication link. Wolski et al. have explained how, for a given quantity of data, the communication bandwidth between nodes in a distributed system affects the performance of the system [39]. Kumar and Lu have also incorporated bandwidth into their energy-based cost model designed specifically for mobile cloud computing [1]. In both cases, communication bandwidth is used as a factor in making computation offloading decisions.

Apart from communication bandwidth, the latency of the link between the mobile device and cloud resource is also an important factor in mobile cloud computing. In their work on interactive user experience on thin clients, Tolia et al. have claimed that latency, rather than bandwidth, is the greatest challenge [40]. Satyanarayanan et al. have pointed out that the long latencies characteristic of Wide Area Networks (WANs) are an obstacle which is unlikely to improve [30]. Due to the geographical distances involved in a global network such as the internet, current network latencies are already of the same order of magnitude as the fundamental lower limit [30]. Satyanarayanan et al. have proposed a possible solution using *"cloudlets"* which are smaller computational resources located in relatively close geographical proximity to the user [30]. This decentralized approach, which resembles the cellular concept in communication networks, eliminates WAN latency by providing localized cloud resources. However, it poses new challenges such as long term data persistence as a user changes location.

As shown in **Figure 3.2**, various external network end-points are secondary sources or destinations of data within the system. These end-points are generally web-based services which support either uploading or downloading of content. The communication between the mobile device and these external end-points would not at first appear to be a primary concern of mobile cloud computing. However, research such as the *"CloudTorrent"* system by Kelenyi and Nurminen has shown that the use of the cloud resource as a communication proxy can provide significant benefit in this situation [41]. Instead of downloading data directly to the mobile device using the BitTorrent protocol, the *"CloudTorrent"* system performs the BitTorrent download in the cloud and transfers the data to the mobile device using Hypertext Transfer Protocol (HTTP) [41]. This use of the cloud resource as a communication proxy has been shown to increase the communication efficiency of this type of system [41].

A further advantage gained by using the cloud as a communication proxy is that data obtained on behalf of the mobile device can also be stored in the cloud for later use. In the general case, in order to perform a computation in the cloud, the mobile device would have to transfer both the input data and the instructions to the cloud. This is subject to the challenges of communication bandwidth and latency as explained above. However, if this data were already cached in the cloud, the mobile device would only have to upload the processing algorithms (instructions) and a reference (pointer) to the data. This would significantly reduce the quantity of data transferred between the cloud and the device. This possibility is highlighted as one of the benefits of cloud computing over the traditional client-server paradigm by Kelenyi and Nurminen [41] and by Kumar and Lu [1].

The computational and communication efficiency of the system could be further increased by using purpose-built cloud resources which already include the algorithms required for processing this data. If both the data and the processing algorithms were already present in the cloud, the mobile device could simply upload references to the appropriate data and the required processing algorithm. This type of system could be realized either by designing mobile applications specifically for this paradigm or by using intermediate subsystems which interact with the mobile applications as well as the cloud resources.

An overview of the principle contributions of this aspect to the overall analysis framework is shown in **Table 3.2**.

Table 3.2: Summary of the Communication Requirements Aspect

Major Considerations: • The quantity of data transferred between the mobile device and its associated cloud resource. • Bandwidth and latency requirements of the communication links. • Communication between the mobile device and external network end-points. • The use of the cloud resource as a communication proxy.
Primary Objective: • Maximize the communication performance and efficiency of the system.

3.4 Mobile Network Impact

By definition, wireless networks are required for communication with all mobile devices. However, mobile computing necessitates the use of wireless networks which support mobility as opposed to nomadic wireless access. In this context, nomadic wireless access means that a user can connect to the wireless network from different geographical locations. *'Mobility'* extends this notion by adding the requirement that the user can maintain an active network connection whilst moving between different geographical locations. This aspect analyses both the impact of the mobile wireless network on the mobile cloud computing system as well as the system's impact on the mobile network.

In a mobile network, connectivity is provided by geographically distributed fixed infrastructure such as cellular base stations or WLAN access points. It is becoming increasingly common for operators to utilize multiple Radio Access Technologies (RATs) to form heterogeneous wireless networks. Since the users of the network are mobile, their position relative to this fixed infrastructure is constantly changing. These physical changes in the wireless communication channel result in a high level of variability in network performance in terms of bandwidth and latency. For example, in a cellular network the user may experience deteriorating signal strength and hence reduced bandwidth as he or she moves further away from the base station. Apart from the variability caused by the mobility of the user, further variability in network performance can be caused by the mobility of other users. In most RATs, the available electromagnetic spectrum and communication bandwidth of a particular access point is shared between all devices using that access point (up to the maximum capacity). As the number of users in a particular area increases, the individual bandwidth allocated to each user in that area will often decrease proportionally. In the extreme case, over-crowding in a particular area could lead to network congestion and disconnection of some users in that area.

Depending on the mobility of the user, the network may perform a handover of this user to a different cell at some point. In the case of a horizontal handover between different cells of the same RAT, the user would generally still be able to maintain an ongoing connection. However, this cannot be guaranteed for all RATs or for vertical handovers between different RATs. Therefore the user may experience temporary unforeseen disconnections from the network as the result of a handover-type event. Similar unforeseen disconnections could also occur if the user moves into an area without coverage (network dead-spot) or if part of the network becomes congested.

Satyanaranan has listed this high variability in the performance and reliability of mobile networks as one of the fundamental constraints of mobility [14]. In mobile cloud computing, it is important to account for this variability in network connectivity. If the system assumes a constant level of connectivity it will not be able to take advantage of periods of increased network performance and so will perform sub-optimally. More importantly, such a system could cease to function when network connectivity falls below the expected level. In the analysis of a mobile cloud computing system, the communication requirements described in the previous section must be evaluated against the variable performance levels provided by the mobile network. Therefore there is an interdependency between this aspect and the communication requirements aspect.

Mobile cloud computing systems can also have an impact on the mobile network. In their discussion of access schemes for mobile cloud computing, Klein et al. have explained that one of the important differences between mobile cloud computing and other networked systems is that the former requires on-demand wireless connectivity and a scalable link bandwidth [42]. In order for mobile cloud computing to be feasible, the Mobile Network Operator (MNO) must be able to provide this type of scalable network connectivity. However, this could lead to increased variability in terms of the utilization of the mobile network. Klein et al. have concluded that concepts such as Intelligent Radio Network Access provide a solution to the challenge of mobile cloud computing in heterogeneous wireless networks [42]. This approach accounts for the status and characteristics of multiple RATs within the network and aims to achieve the best end-to-end performance for users [42].

In dealing with real-world mobile networks, it is also important to analyse the overall available capacity of the network. Whilst the communication technology may be able to meet the communication requirements for a single user, this does not necessarily mean that the network has sufficient scalability to support this technology for all users. Therefore, the capacity of the mobile network must be compared to the communication requirements of the system multiplied by the expected number of users. A more advanced approach would be to perform this comparison on a per-area basis (for example, determining communication requirements per cellular base station). This also confirms the interdependency between this aspect and the communication requirements aspect.

An overview of the principle contributions of this aspect to the overall analysis framework is shown in **Table 3.3**.

Table 3.3: Summary of the Mobile Network Impact Aspect

Major Considerations: • The causes and effects of the high level of variability in network performance in terms of bandwidth and latency. • The possibility of unforeseen temporary disconnection from the network. • The impact of the mobile cloud computing system on the mobile network in terms of scalable bandwidth requirements and overall capacity.
Primary Objectives: • Characterize the impact of the mobile network on a mobile cloud computing system. • Assess the impact of such a system on the mobile network.

3.5 Energy Considerations

Energy considerations are an important aspect in many analyses of mobile cloud computing. As discussed in **Chapter 2**, one of the fundamental challenges in mobile computing is that mobile devices have a finite energy storage capacity [14]. Although this energy is frequently replenished through charging the battery, the objective is to maximize the mobile runtime of the device between charges. As shown by the survey described in **Section 2.3**, this is a very important objective from the users' perspective.

Extending mobile runtime by increasing the energy storage capacity of the battery is generally infeasible due to size constraints. Therefore the energy efficiency of the system must be increased in order to better utilize this limited energy capacity whilst still providing the required functionality. Mobile cloud computing has been proposed as a means of achieving this increase in energy efficiency. This aspect focuses on the way in which energy is consumed in a mobile cloud computing system and how the system can be optimized to maximize energy efficiency.

In their work on the energy aspect of mobile cloud computing, Kumar and Lu [1] as well as Satyanarayanan [43] have proposed various possible approaches to energy conservation on mobile devices. The major themes from these sources are:

- The adoption of a new generation of semiconductor technology in order to shrink the size of transistors and reduce hardware power consumption [1][43].
- The use of stand-by or sleep modes for individual subsystems of the mobile device in order to avoid wasting energy [1].
- The use of energy-aware software to reduce the demands on the mobile device hardware [43].
- Slower execution of computation by operating the mobile device hardware at a lower clock frequency [1].

- The elimination of computation from the mobile device by offloading this computation to a remote resource such as the cloud [1][43].

To some extent, each of these approaches is already being used in mobile computing. At present, mobile devices are generally keeping pace with advances in semiconductor technology. Although the latest technologies may not be adopted immediately due to financial considerations, significant research and development efforts are being undertaken in the area of low-power hardware for mobile devices. The proposal to use stand-by/sleep modes for individual subsystems has already been implemented in the PC domain and is being integrated more fully into the mobile domain. Flinn and Satyanarayanan have demonstrated that energy-aware adaptation for mobile software applications has the potential to yield substantial energy savings [44]. The idea of reducing hardware clock frequencies is being pursued through the design of dual-core and multi-core processors for mobile devices which operate at lower clock frequencies. The final suggestion of eliminating computation from the mobile device can be achieved through mobile cloud computing.

By distributing computation between the mobile device and its associated cloud resource, mobile cloud computing reduces the quantity and possibly also the complexity of computation performed on the mobile device. Since all computation utilizes some energy, this reduction in the quantity or complexity of computation reduces the energy consumption of the mobile device. However, offloading this computation to the cloud requires additional communication between the mobile device and the cloud resource using the mobile wireless network. This communication requires energy which would not have been expended in this way if the computation had been performed on the mobile device.

In order to reduce the device's overall energy consumption using mobile cloud computing, the energy saved by the reduction of on-device computation must exceed the energy expended in communicating with the cloud resource. Flinn et al. have described how the *"Spectra"* system improves energy efficiency by monitoring resources and adapting to changing conditions [35]. In particular, energy-aware software techniques such as goal-directed energy adaptation can be used to achieve the user-specified mobile runtime [35].

Kumar and Lu use an energy-based cost model to determine whether computation should be offloaded to the cloud or performed on-device [1]. In this model, the computational requirements are parameterized as the quantity of computation to be performed (number of operations) and the speed of computation (operations per second). It also parameterizes the communication requirements in terms of the quantity of data to be transferred (number of bits) and the available network bandwidth (bits per second). Since the speed of computation does not change, the outcome of this model depends on the quantity of computation and communication required and the available network bandwidth. This model can be represented graphically as shown in **Figure 3.3** [1].

In **Figure 3.3**, the horizontal axis represents the computational requirements of a specific task (a combination of the quantity and complexity of the required computation) and the vertical axis represents the task's communication requirements (quantity of data to be transferred). The equilibrium lines dividing the sectors show the points at which the energy required for computation is equal to that required for communication. There are two such equilibrium lines because of the variability in factors such as communication bandwidth and cloud efficiency. The upper line corresponds to the best-case scenario (highest available bandwidth) and the lower line to the worst-case scenario (lowest bandwidth). Therefore, in the middle sector, the outcome of the model depends on factors such as network bandwidth and cloud efficiency. The gradients of the equilibrium lines will change depending on the type of device and RAT in use. Lagerspetz and Tarkoma have extended this model by splitting the communication requirements into uplink and downlink requirements since these usually have different energy requirements [45]. They have also used a graphical representation, similar to that shown in **Figure 3.3**, using measurements from a specific device to define the gradients of the equilibrium lines [45].

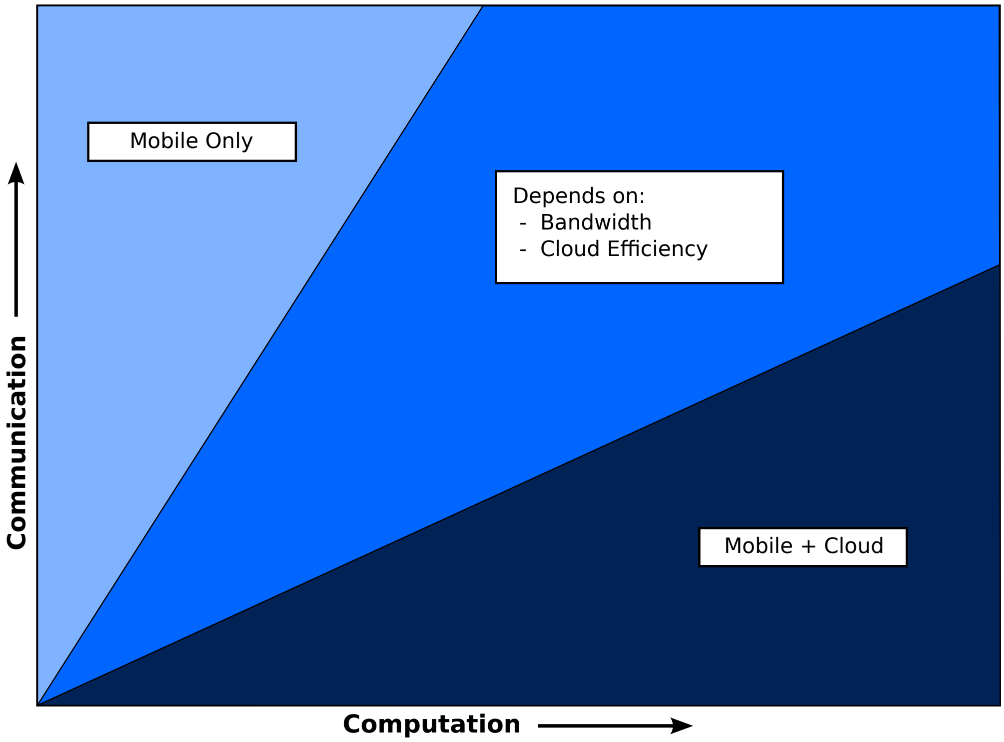

Figure 3.3: A graphical representation of the cost model based on energy consumption adapted from Kumar and Lu [1]. This model is used to determine whether to offload a computational task from the mobile device to the cloud resource. The horizontal axis represents the computational requirements of the task (a combination of the quantity and complexity of the computation) and the vertical axis represents the task's communication requirements (quantity of data to be transferred). This model uses these inputs and various other parameters to minimize the energy consumption of the mobile device.

As explained in the Computational Requirements aspect of this framework, Chun et al. have shown that their *"CloneCloud"* system can decrease the execution time of three test applications: virus scanning, image search and behaviour profiling [2]. They have also shown that this system can decrease the amount of energy required to perform these tasks. In some cases, energy reductions of up to 20 times were measured [2]. The energy requirements were shown to vary depending on the type of application (influencing the quantity and complexity of the computation to be performed) as well as the volume of data to be processed. This means that there is a strong interdependency between the energy considerations aspect and the computational requirements aspect in this analysis framework.

As would be expected, the energy expended for communication increases in proportion to the quantity of data transferred. Energy consumption is dependent on the choice of RAT due to differences between physical layer interfaces. A further result, which is slightly less apparent, is that the bandwidth and latency of the network also influence energy consumption. In their paper on *"Energy efficiency of mobile clients in cloud computing"*, Miettinen and Nurminen have demonstrated that the communication energy per bit decreases as the bandwidth of the network increases [46]. Cuervo et al. have shown experimentally that more energy is required for communication as the latency of the network increases [31]. These results indicate that there are strong interdependencies between this energy considerations aspect and the communication requirements aspect as well as the mobile network impact aspect in this framework.

In the Communication Requirements aspect, the *"CloudTorrent"* system presented by Kelenyi and Nurminen was shown to increase the communication performance of a BitTorrent download [41]. This same test also demonstrated that the use of the cloud as a communication proxy decreased the energy expended by the mobile device in performing this download by more than 60% [41].

As discussed above, mobile cloud computing can result in significant energy savings on mobile devices. Furthermore, since the cloud resource is usually very energy efficient, it may be possible for mobile cloud computing to reduce the total energy consumption of the overall system. In their paper, *"Corollaries to Amdahl's Law for Energy"*, Cho and Melhem have analysed the relationship between parallelization and energy consumption [47]. They have shown analytically that a parallel computing architecture can achieve both a speed-up in execution time and an improvement in energy consumption [47]. Since cloud resources are often based on parallel computing infrastructure, this relationship between parallelization and energy consumption is applicable. Srikantaiah et al. have shown how the energy-aware consolidation of computing can further decrease the energy expended by the cloud [26].

The overall energy consumption of a mobile cloud computing system has been analysed by Chu et al. [48]. By estimating the energy requirements of the external infrastructure, their energy analysis has identified scenarios in which the distribution of computation between the mobile device and the cloud can reduce the overall energy consumption of the whole system [48].

An overview of the principle contributions of this aspect to the overall analysis framework is shown in **Table 3.4**.

Table 3.4: Summary of the Energy Considerations Aspect

Major Considerations: • The reduction in energy consumption on the mobile device obtained through the use of mobile cloud computing. • The optimum distribution of computation between the mobile device and cloud resource in order to maximize energy efficiency. • The possibility of minimizing the total amount of energy consumed by the whole system.
Primary Objectives: • Maximize the energy efficiency of the mobile device. • Minimize the total amount of energy consumed by the system.

3.6 Information Security

Since its origin, it has been accepted that the mobile computing paradigm poses certain risks to information security that are not present to the same extent in a non-mobile context. Compared to non-mobile devices, Satyanarayanan has listed the increased risk of theft, loss or damage of mobile devices as a fundamental challenge of mobile computing [14]. Forman and Zahorjan have argued that since the security of wireless communication can be compromised more easily than that of wired communication, the security implications of wireless communication constitute another of the major challenges to mobile computing [17].

In mobile computing, information security is focused on the security of the mobile device on which the information resides. It is also concerned with securing any communication links over which sensitive information is transferred. Due to the architecture of mobile cloud computing, information security considerations must be extended to cover the important communication link between the mobile device and the cloud as well as the security of information in the cloud. However, existing network and system security technologies can often be used to address the information security considerations of mobile cloud computing in terms of network and cloud security.

In the context of this work a clear distinction is made between the terms *'data'* and *'information'*. Data refers to the binary bits which are processed, stored or transmitted whereas information is a higher level abstraction representing meaningful facts or knowledge. Although information is constituted by data, a set of data taken in isolation may not yield meaningful information. The purpose of this aspect is to ensure the security of information rather than data within the system.

Based on the definition of the term *'Information Security'* provided in the ISO/IEC 27002:2005 standard [49], the major considerations in this field are:

- Authentication (including access control)
- Confidentiality
- Integrity
- Availability
- Accountability
- Non-repudiation

The above concepts are sometimes referred to as *'security services'* and are applicable to both the transfer and storage of information. In mobile cloud computing, authentication (access control), confidentiality and integrity have been identified as important security considerations and are discussed in this aspect. In this context, availability extends beyond information availability to include system availability. Therefore system availability forms a separate aspect of this framework and is analysed in the next section. Accountability and non-repudiation have not featured significantly in any recent literature on mobile cloud computing.

Analysis of mobile cloud computing from an information security perspective must begin with the identification of the type of security services required and the appropriate level of security. These parameters are dependent on the type of information used within the system and so can generally be associated with specific application domains. For example, mobile electronic healthcare (mHealth) applications will almost always involve personal medical information about users. Since this is regarded as sensitive information, the appropriate levels of confidentiality and access control are required. The specific security measures used will depend on how this sensitive information is processed or stored. For example, in the United States of America, when an individual's Protected Health Information (PHI) is transmitted using network applications or electronic systems, these systems must meet the privacy and security requirements of the Health Insurance Portability and Accountability Act of 1996 (HIPAA) [50]. If PHI is moved off the mobile device using a system such as mobile cloud computing, these requirements take effect [50]. However, security requirements such as these can be satisfied using existing technologies as described in the white-paper by cloud provider Amazon on *"Creating HIPAA-Compliant Medical Data Applications with Amazon Web Services"* [51]. Chun et al. have suggested that an alternative approach is to partition the application between the mobile device and cloud in such a way that certain sensitive information never leaves the mobile device [52].

In the mobile computing paradigm, it is assumed that preliminary user authentication is performed by the mobile device in order to grant the user access to the device. User authentication could range from simple password-based techniques through to full biometric authentication. However, user authentication is beyond the scope of this work. Specifically in mobile cloud computing, authentication focuses on the establishment of a trust relationship between the mobile device and cloud resource. Satyanarayanan et al. have highlighted the importance of this trust relationship, especially when dealing with multiple nearby cloudlets [30]. They have proposed that it can be created using either trust-establishment or reputation-based trust [30]. Trust establishment is very straightforward but also computationally intensive because an authentication procedure is performed at the start of each session [30]. Reputation-based trust, which grants trusted status by virtue of the fact that the entity is trusted by others, is more complex but also more efficient in terms of both computation and communication [30]. This shows that there are interdependencies between this information security aspect and the computation and communication requirements aspects since an increase in the level of information security often results in an increase in computational and communication requirements.

Confidentiality ensures that information in the system can only be accessed by authorized entities. In most networked systems including mobile cloud computing, confidentiality is provided by means of encryption techniques. Various encryption algorithms are available and can be used to increase the security of information while it is in transit through the network. However, the cryptographic operations used in these techniques increase the computational requirements of the system which again demonstrates the interdependency between this aspect and the computational requirements aspect.

For information stored on the mobile device, confidentiality can be maintained through the use of storage encryption technologies such as those described in the NIST *"Guide to Storage Encryption Technologies for End User Devices"* [53]. Since the user interacts directly with the mobile device, the user can provide the key to decrypt specific data using password-based or biometric authentication as required. Confidentiality is also a concern for information stored in the cloud especially if public cloud infrastructure is used. Although it is also possible to store data in its encrypted form in the cloud, this negates a significant amount of benefit provided by the cloud because the encrypted information cannot be processed directly [1]. Since the cloud is not necessarily a trusted system, decrypting information in order to process it may result in a security vulnerability. To avoid this problem, Kumar and Lu have suggested the use of information hiding techniques such as steganography in which sensitive information is hidden within a larger dataset [1]. Brenner et al. have presented a method for securing information using homomorphic encryption which allows certain types of processing to be performed directly on the encrypted data [54]. For certain applications, both steganography and homomorphic encryption allow for the processing of information by the cloud without revealing this information to the cloud.

Integrity is an important consideration in the communication and storage of information because it ensures that any modification of the information will be detected. In this context, modification may be caused by either errors in transmission or malicious actions (security attacks). Integrity checking prevents the situation in which modified information is used as the basis for some further operation or decision, leading to an incorrect result or unintended action. However, comprehensive integrity checking is highly computationally intensive and so is often infeasible on certain mobile devices, especially when information is frequently modified. Itani et al. have proposed a solution based on energy-efficient incremental integrity [55]. Their solution requires a cryptographic coprocessor on the cloud system which uses shared-secret collaboration with the mobile device (established during the authentication phase) to generate Message Authentication Codes (MACs) [55]. These MACs can be used to verify the integrity of specific data and can be updated incrementally when this data is modified [55]. Experimental testing of this system showed a 90% reduction in computational requirements compared to a traditional integrity-checking approach [55].

Under certain circumstances, mobile cloud computing can enhance the security of communication between the mobile device and external network end-points. Whenever the mobile device requires a secure connection to an external network end-point, it must establish a trust relationship with this end-point. This is a computationally intensive process since it often involves public key cryptographic techniques and so may be infeasible on certain mobile devices. The mobile device could conceivably interact with several secure network end-points and so would have to establish a separate trust relationship with each of these. By comparison, the mobile device only establishes a single trust relationship with its associated cloud resource.

As shown in **Figure 3.4**, the cloud resource can be used as a secure communication proxy between the mobile device and external network end-points. Using this architecture, the mobile device only has to establish a single secure connection to the cloud resource and then all further secure connections can be established between the cloud resource and the external network end-points. In this way, most of the computationally intensive security operations are performed in the cloud.

For example, digital watermarking is a computationally intensive process used to ensure the authenticity of digital media [56]. By offloading parts of this process from a mobile device to a non-mobile computational resource, Kejariwal et al. have achieved performance increases of more than two orders of magnitude [56].

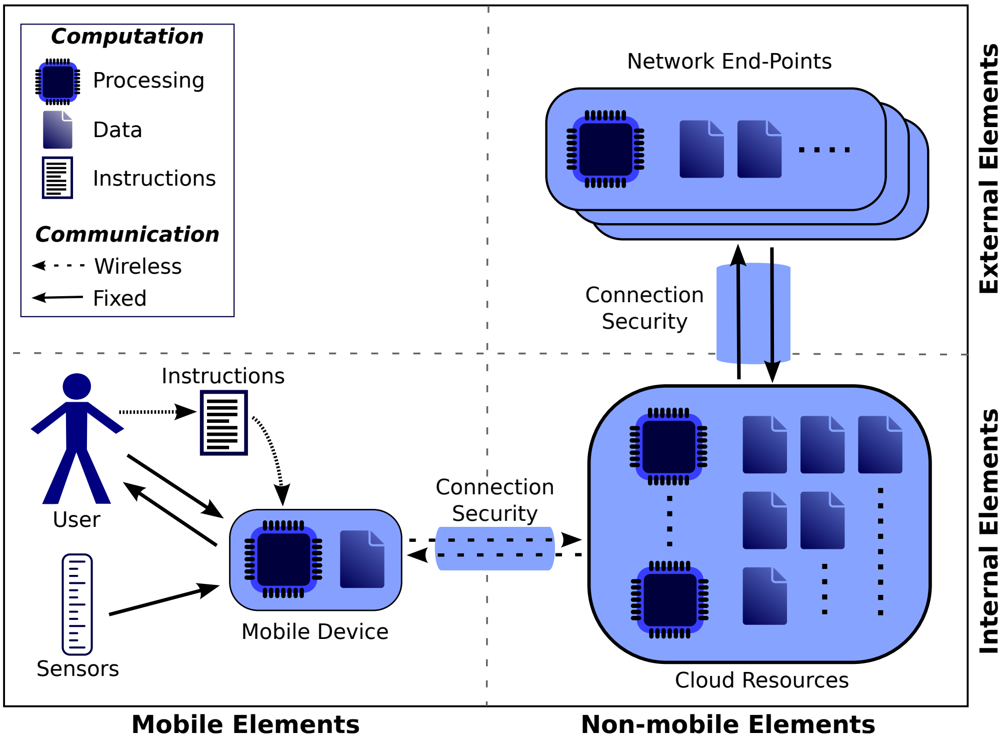

Figure 3.4: A high-level overview showing the use of the cloud resource as a secure communication proxy. Since the connection between the mobile device and the cloud resource remains relatively constant, a single secure link can be established and maintained. The cloud resource can then communicate with various external end-points on behalf of the mobile device and perform the computationally intensive tasks involved with establishing separate secure links to each network end-point as required.

The use of cloud resources has also been suggested as a means to enhance user privacy when using Location Based Services (LBS) through techniques such as spatial cloaking [57]. When using spatial cloaking to interact with an untrusted LBS, the cloud reports the user's geographical location to the LBS as a region containing at least some minimum number of users [57]. Since the untrusted LBS cannot isolate specific users from this group, the user is able to obtain location-specific information without revealing an exact location [57].

An overview of the principle contributions of this aspect to the overall analysis framework is shown in **Table 3.5**.

Table 3.5: Summary of the Information Security Aspect

Major Considerations: • The type of application, the sensitivity of the information involved and the required level of information security. • The use of appropriate technologies to address security challenges and the computational cost of using these technologies. • The use of the cloud resource as a secure communication proxy.
Primary Objectives: • Ensure that the appropriate level of information security is provided by all components of the mobile cloud computing system. • Where possible, to enhance the security of other mobile computing operations through the use of the cloud.

3.7 System Availability

As discussed in the previous section, availability is sometimes considered as one of the major security services. However, in the context of mobile cloud computing, the analysis of system availability and the consequences of unavailability extend beyond information security and so are considered as a distinct aspect in this framework. From a systems engineering perspective, system availability is defined as the time for which a system is operational expressed as a percentage of the total time. In certain literature, this concept is referred to as the reliability of the system but for purposes of clarity the term *'availability'* is used in this work.

System availability is determined by the specific availabilities of the various component subsystems. In the mobile computing paradigm, the availability of the system depends only on the availability of the mobile device since this is a self-contained computing system. In mobile cloud computing, the wireless communication network and the cloud resource are both critical subsystems. Therefore, the availability of a mobile cloud computing system is a combination of the specific availabilities of the mobile device, the wireless communication network and the cloud resource.

In terms of the wireless communication network, unavailability is constituted by the interruption of the communication link between the mobile device and the cloud resource. There are two primary causes of such unavailability. Firstly, as explained in the mobile network impact aspect, an inherent challenge of mobile communication is that the user could move into an area without network coverage (network dead-spot) and so experience unforeseen disconnection from the network. Secondly, the user may experience unforeseen disconnection as a result of a failure in the network. This could be caused by high levels of congestion in a particular area, failure of a radio access subsystem (such as a cellular base station) or failure of one of the critical nodes in the core network. Since any interruption of the connection between the mobile device and the cloud resource could result in unavailability of the mobile cloud computing system, there is an interdependency between the system availability aspect and mobile network impact aspect in this framework. Kumar and Lu have highlighted the importance of system availability through various example scenarios which demonstrate the consequences of unavailability of the wireless communication network [1].

In terms of the availability of the cloud resource, it is important to consider the availability of both the hardware and software subsystems. Failure of either of these subsystems could result in unavailability of the cloud resource. Kumar and Lu have listed some recorded outages of high profile cloud services [1]. Armbrust et al. have cited availability as the number one obstacle in their list of the top ten obstacles to cloud computing [10]. In order to address this challenge, certain public cloud providers offer Service Level Agreements (SLAs) stating a particular level of availability. For example, the Amazon S3 SLA states that the monthly uptime percentage of the service will be at least 99.9% [58].

The obtainable level of system availability limits the types of applications which should utilize this technology. In previous generations of mobile computing, the possible types of applications were constrained by the limited computational capacity of mobile devices. Since these applications provided relatively basic functionality (compared to modern applications), system availability was not a primary concern. However, using the enhanced capacity of mobile cloud computing, it has now become important to consider the availability requirements of mobile applications as well as the consequences of system unavailability. For example, based on current technologies, the use of a mobile cloud computing system as the primary computational tool for emergency response teams (fire fighting or medical emergencies) may not be suitable due to availability considerations. Even though the system may be technically capable of supporting this particular application, the risk and consequences of system unavailability may outweigh the benefit of the system. However, future developments in High-Availability (HA) wireless communication networks and cloud computing technology could address this limitation.

One solution for meeting the availability requirements of a particular application is to ensure that all component subsystems provide sufficient availability guarantees such that the overall system availability meets or exceeds the requirements. It is also possible to increase the system availability using multiple redundant subsystems in order to avoid any single points of failure. At a lower level, several types of wireless communication networks as well as various cloud services already make use of multiple redundancy to increase availability. In certain applications it is possible to mitigate against the consequences of system unavailability through the design of the software systems. For example, the *"Coda"* file system developed by Satyanarayanan et al. provides resiliency to server and network failures in large-scale distributed computing environments [59]. Kistler and Satyanarayanan have explained how the *'disconnected operation'* capability of this file system is particularly important in supporting mobile computing [60].

In terms of information availability, the mobile cloud computing paradigm can provide greater levels of information survivability and availability compared to the previous mobile computing paradigm. In mobile computing, the mobile device is a single point of failure leading to information loss. As explained by Forman and Zahorjan, mobile devices are prone to risks such as physical damage, loss or theft [17]. The use of mobile cloud computing allows information from the mobile device to be replicated in the cloud resource. Once in the cloud, the risk of losing this information decreases significantly due to the multiple redundant storage technologies in use as well as the geographical redundancy benefits of the cloud system.

An overview of the principle contributions of this aspect to the overall analysis framework is shown in **Table 3.6**.

Table 3.6: Summary of the System Availability Aspect

Major Considerations: • The possible causes of system unavailability. • The type of application and the level of system availability it requires. • The use of availability guarantees and multiple redundant subsystems to meet availability requirements.
Primary Objective: • Ensure that the system meets the availability requirements of applications using this technology.

3.8 Application Usability

As discussed in **Chapter 2**, the constraints on the physical size of mobile devices are a fundamental limitation of mobility. This small physical size generally limits the User Interface (UI) capability of the device as described by Forman and Zahorjan [17]. From a usability perspective, the effectiveness of mobile device UI subsystems such as graphical displays and keyboards/keypads is often dependent on the available physical surface area.

UI constraints can limit the quantity of information exchanged between the user and mobile device as well as the rate at which this transfer takes place. For example, the size and resolution of the graphical display limits the amount of data which can be shown to the user and the relatively small mobile keyboard restricts the rate at which the user can enter information. Apart from the hardware limitations, Chittaro has pointed out that the highly variable physical environment in which mobile devices are used introduces further complications [61]. For example, as a consequence of mobility, the lighting conditions may range from glare to total darkness [61]. These UI constraints affect the type of applications which should be provided through the mobile computing paradigm as well as the level of functionality of these applications.

Before the advent of mobile cloud computing, the limited computational capacity of mobile devices acted as a moderating factor in this regard. It was often infeasible to compute quantities of useful information that would exceed the UI capability of a mobile device. With the augmented computational capacity endowed by mobile cloud computing, it is now possible to run many new applications on mobile devices. However, since this technology does not increase the hardware UI capability, it is significantly less difficult for applications to exceed the UI constraints of a mobile device. Therefore it is important to consider application usability in the analysis of mobile cloud computing systems.

Due to these mobile UI constraints, application usability places an upper bound on the computational and communication requirements of an application. For example, in a multimedia or gaming application it is not beneficial to render graphical output at a higher resolution than that of the graphical display on the target mobile device. Similarly, this information need not be transferred to the device at a higher rate than it can be displayed (except in the case of buffering). This aspect therefore exhibits interdependencies with the computational and communication requirements aspects.

However, developments in mobile UI hardware are continually reducing the impact of these UI constraints. As explained in **Chapter 2** these developments include technologies such as multi-purpose touch screens which allow the same physical area to be used for both user input and output (or even a simultaneous combination thereof) as required. Mobile cloud computing can also be used to enhance non-hardware UI capabilities such as handwriting recognition or voice recognition. Although these are not new techniques, their adoption in previous mobile computing systems was limited by the substantial data processing and storage demands they place on the mobile device [17]. Visualization of information on a mobile device is also an active research area. Chittaro has listed some of the recent advances in the visualization of different types of information on mobile devices [61]. Lu et al. have proposed that the rendering of graphical output can be offloaded to the cloud in order to enhance the mobile computing user experience [62]. In this proposal the relatively small size of a mobile screen is actually advantageous since it limits the amount of data transfer required [62].

In certain applications, the use of peripherals can overcome the UI constraints of a mobile device. Input peripherals such as cameras, GPS receivers, accelerometers and NFC transceivers can supply a greater quantity of information at a significantly higher rate than traditional user input. The peripherals themselves can be either fixed or mobile systems. For example, recent hardware developments have made it possible to display graphical output from the mobile device via large high-resolution fixed external displays or personal mobile projectors [63].

Terrenghi et al. have described a scenario in which the mobile device becomes the control interface to various external peripherals such as large fixed displays or interactive surfaces [64]. They have suggested that in the future, these peripherals will become widely distributed throughout the environment (in both public and private locations) and will be made accessible to mobile devices [64]. They have proposed that by connecting a mobile device to these distributed peripherals, the usability of the system can be enhanced whilst still maintaining a high level of mobility [64]. Although the use of peripherals is often dependent on the application, in cases where they are used, the UI capability of the mobile device is substantially enhanced.

An overview of the principle contributions of this aspect to the overall analysis framework is shown in **Table 3.7**.

Table 3.7: Summary of the Application Usability Aspect

Major Considerations: • The UI capability and UI constraints of the mobile device. • The effects and impact of these constraints on the functionality of mobile applications. • The possibility of augmenting the device's UI capability using software UI techniques or external peripherals.
Primary Objectives: • Ensure that the impact of mobile device UI constraints is taken into account. • Improve mobile UI capabilities using cloud-enhanced UI techniques.

3.9 Conclusion

The theoretical analysis framework defined in this chapter is a structured consolidation of the various considerations relevant to mobile cloud computing identified in recent scientific literature. This framework has been designed to facilitate the standardized analysis and comparison of mobile cloud computing systems.

Three important requirements which guided the design of this analysis framework are that it should be comprehensive in terms of its coverage, flexible in its application and enduring in its relevance. The high-level systems approach used in this chapter allows for the inclusion of a broad spectrum of analysis criteria and so ensures the comprehensive nature of this framework. This approach also results in a flexible and enduring framework since it does not rely on the use of low-level quantitative metrics and benchmarks. Although useful, low-level metrics are inherently linked to a specific purpose and so limit the flexibility of the framework. Furthermore, due to the rapid ongoing development of this technology, low-level metrics must be constantly updated to remain current with the latest generation of measurement techniques and benchmarks. In order to remain relevant, this framework is primarily based on the fundamental concepts behind mobile cloud computing which are largely independent of any particular technology.

The various considerations in this framework are grouped into seven logical aspects in order to provide the required degree of structure in the analysis. The major considerations of each aspect are summarized at the end of the relevant section. Based on these seven aspects, the primary objectives of this framework are as follows:

- Maximize the computational performance of the mobile device.
- Maximize the communication performance and efficiency of the system.
- Characterize the impact of the mobile network on a mobile cloud computing system and assess the impact of such a system on the network.
- Maximize the energy efficiency of the mobile device and minimize the total amount of energy consumed by the whole system.
- Ensure that the appropriate level of information security is provided and, where possible, enhance this security through the use of the cloud.
- Ensure that the system meets the availability requirements of applications using this technology.
- Ensure that the impact of mobile device UI constraints is taken into account and improve UI capabilities using cloud-enhanced UI techniques.

As a result of the grouping of considerations into aspects, several interdependencies between aspects have been identified. Each of these relationships indicates that the relevant aspects have a direct effect on one another. For example, there is an interdependency between the information security aspect and the computational requirements aspect. As explained in **Section 3.4**, applications involving sensitive information require additional security measures such as data encryption which increase the computational requirements of the system. By contrast, the computational requirements aspect is orthogonal to the system availability aspect because the availability requirements of a particular system neither increase nor decrease the computational requirements of that system.

These interdependencies between aspects are derived from the major considerations identified in each aspect. When a consideration in one aspect depends on or modifies a variable which is part of another aspect, an interdependency exists between these two aspects. Interdependencies may be unidirectional or bidirectional and, in the case of quantitative considerations, may represent either direct or inverse proportionality. The precise nature of each interdependency is explained in the relevant section of the framework. An overview of all the interdependencies which have been identified in this framework is shown in **Table 3.8**.

Table 3.8: Interdependencies Between Aspects in the Framework

	Computational	**Communication**	**Mobile Network**	**Energy**	**Security**	**Availability**	**Usability**
Computational	—	•		•	•		•
Communication	•	—	•	•	•		•
Mobile Network		•	—	•		•	
Energy	•	•	•	—			
Security	•	•			—	•	
Availability			•		•	—	
Usability	•	•					—

When using the analysis framework, the interdependencies for a particular aspect denote the set of other aspects which could be affected by the considerations in that aspect. It is important to take all these interdependencies into account. If the aspects are used in isolation, there is a risk that the effects of one aspect may not be correctly reflected elsewhere. Therefore, these interdependencies serve to complement the main aspects of the analysis framework.

Furthermore, these interdependencies between aspects demonstrate the importance of performing a comprehensive analysis. As shown in **Table 3.8**, all aspects of the framework have at least one interdependency. This means that it is not possible to omit any aspects of the framework in a particular analysis. Therefore, these interdependencies also serve as a form of validation for the analysis framework.

By standardizing the analysis criteria, this framework facilitates comparisons between different systems based on the concept of mobile cloud computing. This is demonstrated in **Chapter 4** where the framework is used to analyse and compare various mobile application domains in terms of their suitability for mobile cloud computing and the potential benefit which they could derive from this technology. Various major considerations identified in this framework are also used in **Chapter 5** in the analysis and comparison of different approaches to the implementation of mobile cloud computing systems.

Therefore, the theoretical analysis framework defined in this chapter fulfils the second objective of this research as specified in **Chapter 1**.

Chapter 4

Enhanced Mobile Application Domains

4.1 Introduction

As explained in **Chapter 2**, there is currently a trend towards the use of mobile computing as indicated by the increasing numbers of MCCDs as well as mobile apps. In this context, the term *'mobile app'* refers to any software component which runs on the MCCD and provides definable functionalities. This terminology is in line with that frequently used in the context of smartphones (currently the most pervasive type of MCCD) in which all executable software packages are referred to as apps. Mobile apps can provide functionality directly to the user or to other apps on the device. Certain apps form the built-in functionality of modern mobile operating systems and are shipped with the mobile device. Others are available for download, often through centralized repositories known as mobile *'app stores'*.

The purpose of this chapter is to demonstrate the use of the framework defined in **Chapter 3** in the analysis of various mobile application domains. This is important for three main reasons: Firstly, the successful use of the framework serves as a form of validation of the framework itself. Secondly, the analysis of specific application domains results in useful insight and conclusions relevant to each domain as presented in the respective sections. Thirdly, this chapter demonstrates the process of using the analysis framework. Based on the analysis of the specific application domains, the process of using this framework is generalized so that it can be applied to other mobile cloud computing systems as explained in **Section 4.8**.

In order to maintain a sufficient level of flexibility, this research takes a high-level fundamental view of mobile computing. Instead of directly analysing specific mobile apps, this chapter presents the analysis of various mobile application domains. A mobile application domain is an abstract construct representing a specific purpose (application) for which mobile devices can be used as computing tools. A mobile app can be classified into a particular application domain based on the functionality provided by the app. Each application domain may consist of any number of relevant mobile apps.

Since many apps provide similar functionality, it is more useful to analyse the application domain as a whole and use specific mobile apps for illustrative purposes where necessary. This allows the insight and conclusions of the analysis to be generalized for all apps (including future apps) within that domain.

This approach also eliminates ambiguity in the analysis of apps which can be used for multiple purposes. A single app may exhibit different requirements or characteristics depending on the purpose for which it is used. Direct analysis of this app would therefore be difficult because of its multi-purpose nature. Since an application domain is defined using functional characteristics, it ensures that all constituent apps will be relatively homogeneous from a technical perspective. In the case of a multifunctional app, the different purposes of the app can be successfully analysed in their respective application domains. For example, a productivity suite app could provide mobile scientific computing functionality (spreadsheets and databases) as analysed in **Section 4.2** as well as multimedia editing capabilities as discussed in **Section 4.6**.

In recent scientific literature, a number of mobile application domains have been identified as good candidates to benefit from mobile cloud computing. There are also various commercial systems which currently use this technology. The domains analysed in this chapter have been sourced from both recent scientific literature as well as commercial endeavours. The in-depth technical details of commercial systems are generally not available but the documented high-level functionality is sufficient for this comparative analysis. These domains have been selected based on the degree of benefit that each domain can derive from the use of mobile cloud computing. Although this chapter features some of the most important mobile application domains, this list cannot be considered exhaustive since new domains are constantly being developed.

Although some of these domains are discussed in recent scientific literature, they are only analysed in terms of a smaller subset of criteria. For each application domain, the comprehensive analysis in this chapter aims to determine which aspects will benefit from the use of mobile cloud computing and to provide a qualitative assessment of the degree of benefit. This process also highlights some important issues in each domain which need to be addressed in order to ensure that all applications in that domain are suitable for mobile cloud computing. It must be noted that some aspects of particular application domains have received very little attention in any scientific literature. However, these aspects have still been analysed based on the known requirements and the inferred characteristics of the application domain.

Using the above rationale, the following six application domains have been selected for analysis in the indicated sections of this chapter:

- Mobile scientific computing and simulation - **Section 4.2**
- Mobile electronic healthcare services (mHealth) - **Section 4.3**
- Mobile tools for education and training (mLearning) - **Section 4.4**
- Advanced mobile Human-Computer Interaction (HCI) - **Section 4.5**

- Entertainment and multimedia services - **Section 4.6**
- Mobile gaming applications - **Section 4.7**

In each of the sections listed above, the following structure is used: The section begins with an introduction to the application domain and its high level functionality. Following the introduction, specific examples of noteworthy research and commercial endeavours from that domain are discussed in order to highlight elements that are relevant to the analysis. The framework defined in **Chapter 3** is then used to facilitate a structured analysis of the domain. For each aspect of the framework, specific examples are used to show the extent to which each of the major considerations of that aspect is manifest in the application domain. Each section concludes with a summary of the application domain and the important results of the analysis.

By using this structured framework in the analysis of these application domains, it is possible to make useful comparisons between different domains. However, since this analysis is mostly qualitative in nature, these comparisons are not made in absolute terms. Therefore, in this chapter, when characteristics are assigned a *relative* qualitative value, this should be interpreted as being relative to the other application domains. Where required, a baseline for these relative comparisons can be represented by the capacities and capabilities of a current-generation MCCD which has not been enhanced using cloud resources.

Some of the conclusions are specific to a particular application domain and these are presented at the end of the relevant section. The overall conclusions and recommendations of the analyses described in this chapter are presented in **Section 4.9**.

4.2 Mobile Scientific Computing

Mobile scientific computing refers to the use of mobile devices for any form of scientific or technical computing. This includes techniques such as modelling or simulation in order to obtain solutions to scientific problems. The usefulness of any scientific computing process is generally influenced by the quantity and accuracy of the input data as well as the computational capacity of the system on which it is performed. Mobile cloud computing aims to enhance this application domain by augmenting the computational capacity of mobile devices.

An example of this application domain is a system for drought prediction using mobile devices which has been discussed by Masinde and Bagula [65]. They have motivated the importance of this work by explaining how drought affects more people than any other natural hazard but its effects can be mitigated if accurate and timely predictions are available [65]. The advantage of performing this computation on mobile devices as opposed to PCs is that, in developing countries, mobile devices are far more prevalent than PCs thus making this type of scientific computing broadly available [65]. Furthermore, the mobile devices can perform a dual role by both collecting input measurements and running simulations. The challenge in this domain is that these applications generally require greater computational capacity than is available on a single mobile device.

An example from the commercial domain is the *"Wolfram Alpha for Mobile"* app from Wolfram Research [66]. As shown in **Figure 4.1**, this app provides a mobile interface to the cloud-based mathematical, technical and scientific computation system developed by the same company [66]. By removing the dependence on PCs, this app allows users to access this cloud-based functionality whilst mobile. However, as in the previous example, the required algorithms are too complex to run on a mobile device and so are performed in the cloud.

Scientific computing applications usually satisfy their high computational

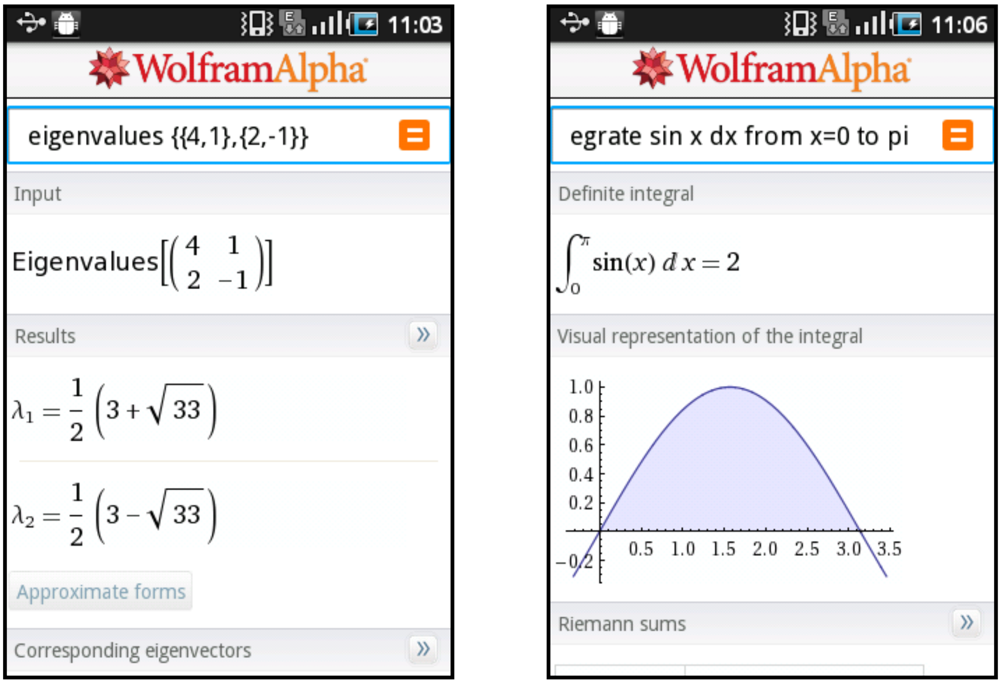

Figure 4.1: Screenshots of the *"Wolfram Alpha for Mobile"* app from Wolfram Research [66]. This mobile scientific computing system uses cloud resources to interpret the user's input and compute the required output.

requirements through the use of High Performance Computing (HPC) infrastructure. Some existing cloud services can already provide the required computational capacity and have been specifically designed for scientific computing. For example, Amazon has recently introduced the *"Cluster"* family of instance types (types of cloud resources) which have been optimized for HPC applications [67]. This family includes the *"Cluster Compute"* and *"Cluster GPU"* instance types, the latter providing highly parallel performance usually obtained through a cluster of Graphics Processing Units (GPUs) [67]. Within the cloud-based HPC cluster, existing technologies such as *"MapReduce"* can be used to distribute tasks amongst all the computational nodes [68]. The use of *"MapReduce"* technology in the context of mobile computing has been discussed by Marinelli in his work on the *"Hyrax"* system [3].

Applying the analysis framework defined in the previous chapter to this application domain yields the following outcomes:

Computational Requirements: In this domain, applications are characterized by very high computational requirements because of the large quantity of computation required and the relatively high complexity of these operations. Often, a significant amount of data will also need to be processed. These requirements usually exceed the computational capacity of a single mobile device. Masinde et al. have proposed that the drought prediction system described above could be run on a grid of mobile phones [69]. Although mobile grid computing differs from mobile cloud computing as explained in **Section 5.2**, both share the objective of augmenting the computational capacity of a mobile device. As demonstrated by the *"Wolfram Alpha for Mobile"* system [66], mobile devices can utilize the significant computational capacity provided by the cloud to support the computational requirements of mobile scientific computing applications.

Communication Requirements: The amount of data transferred between the device and the cloud depends on the type of scientific computing performed. From the above examples, the drought prediction system would transfer a larger amount of data than the mathematical computation system. However, the transfer of this data is not subject to time constraints in this domain since the accuracy of the result is more important than the time required for computation. This means that the bandwidth and latency requirements and thus the overall communication requirements are relatively low in this domain.

Mobile Network Impact: Since the communication bandwidth and latency are not critical factors, this type of application is feasible over a mobile network. The input data can be transferred to the cloud as it becomes available using a low bandwidth link which would not place a significant load on the mobile network. Since this domain may involve long-running processes, it will be necessary to adequately address the possibility of unforeseen disconnection between the mobile device and the cloud.

Energy Considerations: Since this type of computation is generally not possible on a single mobile device, the decision to offload computation to the cloud does not involve the energy cost model. Therefore, energy is not a primary consideration in this domain. However, the mobile cloud computing approach is more energy efficient than the mobile grid alternative because significantly less energy is required to perform a computation on the cloud infrastructure than on a group of mobile devices.

Information Security: In this domain, security is a moderate concern and is also dependent on the specific information involved. The objective of using mobile devices for this type of application is to broaden access to information, especially in developing areas. Based on this objective, it is likely that the results of mobile scientific computing would be shared in an open manner. However, since original data is being generated, some level of security will be required both while the information is being transferred and while it is stored in the cloud. This can be achieved using existing security technologies.

System Availability: Since this type of computation was not previously possible from mobile devices, there are no strict availability requirements. Especially in rural areas, users should be made aware of the possibility that the system may be unavailable due to lack of network connectivity. However, the impact of unavailability is relatively low for the examples described above.

Application Usability: The UI capabilities of the mobile device will generally limit the amount of information involved in mobile scientific computing applications. For example, the device's screen size will constrain the amount of detail in graphical plots or figures generated by either of the systems described above. This limitation may require numerical data to be condensed or aggregated in order to be accurately displayed. The mobile device UI also limits the rate of input from the user to the device. However, the use of peripherals such as input sensors or output displays may remove these limitations in specific applications.

Overall, mobile cloud computing is the critical enabling technology for mobile scientific computing because of the high computational requirements in this domain. Communication requirements are relatively low because bandwidth and latency are not critical thus making this type of application feasible over a mobile network. Since mobile scientific computing would not be possible without the augmented computational capacity provided by the cloud resource, energy is not a primary consideration in this domain. Security and availability are dependent on the specific application but are not critical factors. Usability considerations are very important because it is likely that they will limit the quantity and rate of information transferred between the system and the user.

4.3 Mobile Electronic Healthcare Services

The mobile electronic healthcare services (mHealth) domain includes all medical and healthcare applications which are enhanced through the use of mobile devices. The use of mobile devices for healthcare purposes has been discussed substantially in scientific literature. Hameed has explained that as society becomes more mobile, the demand for mobile healthcare service will increase [70]. Varshney has listed several use-cases within this domain including mobile telemedicine, personalized monitoring and location-based medical services [71]. All these use-cases benefit from the pervasive nature of mobile devices which can be used to extend the reach of healthcare services.

Applications in this domain often obtain input data through physiological sensors connected to the mobile device. In the *"CellScope"* project, Breslauer et al. have designed a high-resolution microscope attachment for camera-enabled mobile phones as shown in **Figure 4.2** [72]. Using this attachment, they have demonstrated that mobile devices can play an important role in telemedicine applications [72]. Various other mobile health-related sensors, including blood pressure sensors and heart rate monitors, are currently under development or are already available. These sensors may either have physical connections to the mobile device or utilize wireless Personal Area Network (PAN) technologies such as Bluetooth.

Nkosi and Mekuria have explained how mobile healthcare applications can be enhanced through the use of mobile cloud computing [73]. They have

Figure 4.2: Photograph of the microscope attachment for mobile devices developed by the *"CellScope"* project [72]. This system uses the camera of the mobile device to capture images at high magnification and is thus useful in mobile healthcare applications [72]. Processing of the resulting image can be augmented through the use of mobile cloud computing.
(Image adapted from Breslauer et al. [72]).

described various computationally intensive operations used by mHealth applications which can be offloaded to the cloud resource [73]. These operations include multimedia signal processing and digital filtering required to extract information from the data obtained through the physiological sensor [73]. Nkosi and Mekuria have also identified the critical importance of securing these mHealth applications and have discussed the security considerations of such a system [73].

In their work on the *"Mobile Cloud for Assistive Healthcare (MoCAsH)"* system, Hoang and Chen have provided a comprehensive analysis of the use of mobile cloud computing in mHealth and have discussed how the cloud computing paradigm has addressed the limitations of the grid computing paradigm in this context [74].

Using the analysis framework, the major outcomes and requirements in the mHealth application domain are:

Computational Requirements: mHealth applications generally have relatively high computational requirements due to the signal processing involved in extracting information from the sensor data as explained by Nkosi and Mekuria [73]. The quantity of data generated by these sensors can also become relatively large thus further increasing the application's computational requirements. In some cases, these computational requirements exceed the capacity of the mobile device thus necessitating the use of the cloud resource. In cases where the use of the cloud is optional, offloading computationally intensive tasks may improve the computational performance of the system.

Communication Requirements: Similarly to the mobile scientific computing domain, certain use-cases in the mHealth domain may require the transfer of relatively large amounts of data from the mobile device to the cloud. Since these are not real-time systems, bandwidth and latency are not critical factors and the communication requirements are relatively low. In terms of communication efficiency, the cloud can be used as a means of transferring medical information from the mobile device to authorized healthcare professionals for further analysis. Doukas et al. have demonstrated a prototype implementation of a mobile healthcare information management system which enables the management and communication of patient health records (including medical images) using mobile cloud computing [75].

Mobile Network Impact: From a mobile network perspective, the low communication requirements of this domain make it feasible over most current mobile RATs. Although infrequently required, large data transfers can be accomplished using low-bandwidth channels. This domain will not have a significant impact on the mobile network. Although it is unlikely that unforeseen network disconnection will have a significant impact on applications in this domain, this possibility should be taken into account in the design of the communication protocols for these systems.

Energy Considerations: Energy can be used as a factor in determining the optimum distribution of computation between the mobile device and the cloud resource. Certain operations may still be too complex to perform on the mobile device whilst others could be performed either on the device or in the cloud. In cases where offloading is optional, the energy cost model can be used to maximize the energy efficiency of the mobile device by comparing the computational and communication requirements based on their respective energy costs. If this type of application is used in rural areas, the importance of minimizing energy consumption on the mobile devices increases significantly.

Information Security: Since this domain deals with sensitive personal medical information, security is a critical requirement. Using mobile cloud computing, sensitive information may potentially be transferred over a non-private network and stored on public cloud infrastructure. This necessitates the use of additional security measures to enforce the appropriate levels of confidentiality, integrity and access control. Legislation such as the Health Insurance Portability and Accountability Act of 1996 [50] specifies the type of information security which must be provided for network transfer and cloud storage. These additional security measures, such as encryption and integrity checking, will require additional processing of any data leaving the mobile device thus increasing the computational requirements of the system.

System Availability: The availability of systems in this domain could potentially become a critical requirement if users become dependent on a particular mHealth service. This would be especially true in rural areas where healthcare workers visit remote settlements on an infrequent basis. Therefore, it is important to ensure that all subsystems can provide the required availability guarantees to ensure that the overall system availability requirements are achieved.

Application Usability: Although this domain may involve large quantities of data, application usability is not a critical factor. On the input side, the data is obtained through specialized physiological sensors which provide the required quantity of data at the correct rate. Since the users are unlikely to be medical professionals, the processing of this data yields a highly condensed and simplified result. For example, the most likely output would be a message informing the user that there is no problem or advising the user to seek consultation with a medical professional. Therefore, the UI constraints of the mobile device do not limit either the input or output of information in this application domain.

The high computational requirements and relatively low communication requirements of this domain make it a very good candidate to benefit from mobile cloud computing. Since they are feasible over most mobile networks, these types of applications can significantly increase the reach of healthcare services. The cloud resource can be used to improve communication efficiency by facilitating the transfer of information between the patient and an authorized healthcare professional. There is potential to reduce energy consumption by offloading certain operations to the cloud. Information security is critical in this domain and additional security measures are required for data transfer and cloud storage. The appropriate level of system availability must be ensured since users may become dependent on the system. The use of physiological sensors removes the UI constraints of mobile devices in this application domain.

4.4 Mobile Tools for Education and Training

The use of MCCDs as educational tools is an extension of the electronic learning (eLearning) paradigm and is termed *'mLearning'*. Mobility allows learning to take place outside the classroom utilizing mobile devices to provide context-relevant educational information. The connected nature of these devices also facilitates educational collaboration between students.

Farooq et al. have highlighted various advantages of mLearning by presenting a comprehensive example scenario of the use of this type of application [76]. In this scenario, some students are performing an educational experiment whilst in the field [76]. They use mobile devices to compare the results to previous data and to communicate the new data to their colleagues at another location [76]. This scenario assumes that the relevant educational content can be provided via the mobile devices and that a collaborative platform exists for sharing this content. Content and collaboration are the two fundamental considerations in the mLearning paradigm and both can benefit significantly from the use of mobile cloud computing.

Currently various open educational resources such as virtual text books and multimedia presentations are available online. For example, the popular video sharing web site, *YouTube*, provides a category of resources specifically for education [77]. It is often the case that these resources are designed for use on PCs rather than on mobile devices. Due to the wide variety of possible MCCDs it is infeasible to redesign and maintain these resources for every possible device configuration. In the mobile cloud computing paradigm, the cloud resource can be used as a communication proxy to retrieve content from external network end-points and dynamically convert it to match the computational capacity and UI capability of the mobile device. For example, instead of downloading a large presentation directly to the device, the cloud resource could download the presentation, reformat and optimize it specifically for that particular type of device and then transfer only the relevant data to the mobile device over the wireless network. At the Nelson Mandela Metropolitan University, a system of this nature is being developed in which cloud resources (Amazon EC2 [24] and Amazon S3 [37]) are used to reformat mathematics content, which was originally designed for PC users, and deliver it to resource-constrained mobile devices [78].

Other examples of mobile tools for education and training are shown in the screenshots in **Figure 4.3**. This figure shows a mobile application for streaming educational video content and a mobile interface to a university's library search system.

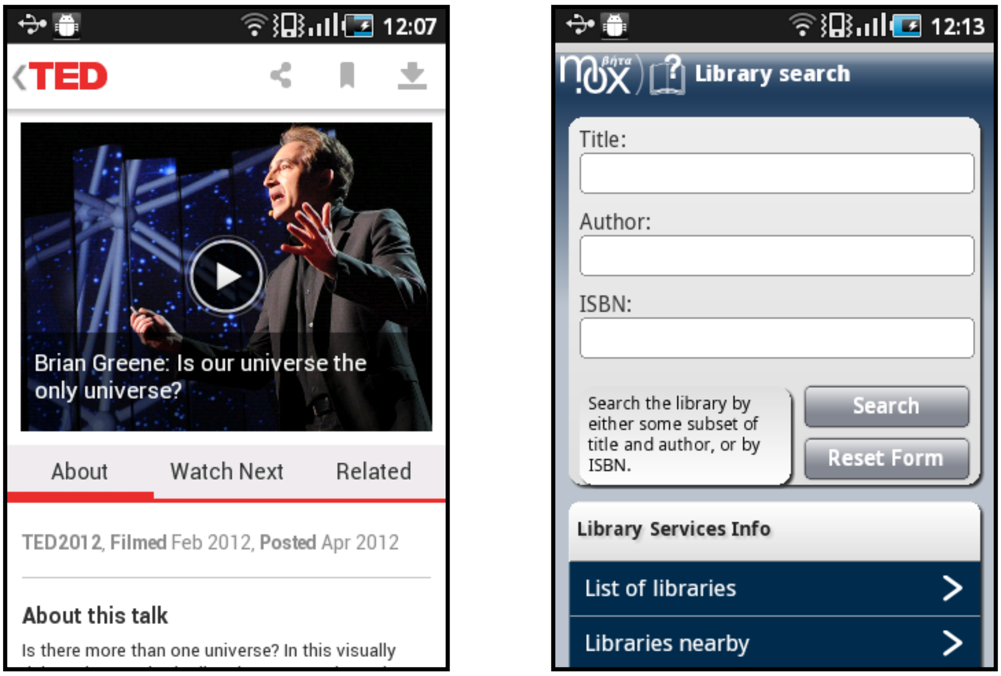

Figure 4.3: Screenshots of mobile tools for education and training. On the left, the *"TED"* mobile app *(www.ted.com)* streams educational videos from the cloud to a mobile device. On the right, the *"Mobile Oxford"* application provides tools such as library search functionality via a mobile device.

Since multimedia educational resources usually involve a significant amount of data, the constrained data storage capacity of the mobile device limits the amount of content which can be stored directly on the device. Cloud resources can be used to augment this limited storage capacity. Mobile cloud computing can also provide a collaborative environment in which users can interact with each other and share information via the cloud. Each MCCD still communicates with its associated cloud resource and the collaboration is facilitated internally in the cloud.

The benefits of mLearning are even greater in developing countries where access to mobile devices far exceeds access to PCs. Masud et al. have described an mLearning system designed for use in Bangladesh which utilizes cloud resources to provide some of the functionality described above [79].

In one of their example applications, Tanwer et al. have combined the augmented capacity provided by mobile cloud computing with the data obtained from peripheral sensors on the mobile device in the context of mLearning [80]. In this example, mobile cloud computing is used to augment a virtual learning scenario based on real-world contextual data obtained from the mobile device [80].

Application of the analysis framework to the mLearning paradigm highlights the following points:

Computational Requirements: Compared to the previous domains, the computational requirements in this domain are moderate to low. The processing requirements involve the reformatting of content to match the capabilities of the mobile device. Since data storage forms part of this aspect, there is still a requirement to store educational content in the cloud.

Communication Requirements: Communication requirements are important in this domain, specifically the transfer of content from the cloud to the mobile device as well as the collaboration between learners via the cloud. Although not critical, the bandwidth and latency of the communication links do have an effect on the quality of the output and the user experience. The use of the cloud resource as a communication proxy can significantly enhance the communication efficiency of applications in this domain.

Mobile Network Impact: It is possible to achieve the same mLearning objectives without using mobile cloud computing. However, the use of this technology is beneficial from a network perspective because it improves the communication efficiency of the system and thus reduces the load on the mobile network. In order for this system to be used effectively in rural areas, it is important to ensure that the required level of network coverage is available.

Energy Considerations: From an energy perspective, the use of the cloud resource improves the efficiency of mLearning applications. By downloading only the required data from the cloud, the mobile device expends less energy on communication. The processing and rendering of the content on the mobile device also requires less energy because of the conversion and optimization performed by the cloud resource.

Information Security: Similarly to the mobile scientific computing domain, the objective of this domain is to broaden the reach of mLearning. Since most of the educational content is open and collaboration between learners is encouraged, information security is not a critical aspect in this domain. A possible benefit of mobile cloud computing is that collaboration and communication between learners can be automatically moderated by the cloud resource.

System Availability: Although not critical, system availability will have an impact on the success of this technology in broadening access to mLearning. However, obtaining a higher level of system availability usually requires additional financial resources which could have been used to extend the reach or enhance the functionality of the system. Therefore system availability must be balanced with improvements in other areas of this domain.

Application Usability: Usability is a critical aspect of mLearning. In this domain UI considerations are asymmetric because the output of information from the device to the user is significantly more prevalent than information input. The cloud resource plays a pivotal role in reformatting and optimizing content for output on mobile devices. Glavinic et al. have presented a holistic approach to enhancing usability in the design of mLearning applications [81]. Although the limited UI capabilities of MCCDs are not ideal, these devices can still provide beneficial mLearning services through the use of mobile cloud computing.

In this domain, the two fundamental considerations of content and collaboration can both benefit from the use of mobile cloud computing. The cloud resource provides the capacity to store educational content on behalf of the mobile device and can be used to reformat and optimize this content. The use of the cloud as a communication proxy improves the communication efficiency of the system and thus reduces the load on the mobile network. This increases energy efficiency and extends the mobile runtime of the device which is a particularly important consideration in rural areas. Information security is not a primary concern due to the open nature of this domain. Availability considerations, although important, should be balanced with other objectives. Usability is a critical aspect in this domain but can be improved through the use of mobile cloud computing.

4.5 Advanced Human-Computer Interaction

This application domain includes all systems which facilitate interaction or transfer of data between the mobile device and the user. Research in the field of Human-Computer Interaction (HCI) is well established in the PC domain and is increasing in the mobile domain. Since the hardware capabilities of mobile devices are still evolving relatively rapidly, HCI for mobile devices (mobile HCI) is still an area of ongoing research.

The constraints on the physical size of a mobile device limit the available surface area and thus the hardware UI capabilities of the device. Providing useful user interaction within the constraints of this physically small UI has been identified as one of the fundamental challenges of mobile computing by Forman and Zahorjan [17]. New methods in mobile UI design are also starting to emerge such as the constraint-based UI design method for mobile computing devices proposed by Niu et al. [82].

In an ongoing effort to enhance mobile HCI, various techniques can be used to facilitate user interaction. The most common of these are speech recognition (voice control), handwriting recognition and gesture commands. Although these are already feasible on certain mobile devices, the level of accuracy in each of these techniques is directly related to the computational requirements of the system due to the complexity of the computation involved. In the mobile cloud computing paradigm, the augmented computational capacity provided by the cloud resource can be used to enhance the accuracy and efficiency of these UI techniques as well as enabling new forms of advanced HCI for mobile devices. **Figure 4.4** shows screenshots of two mobile apps which use the augmented capacity of the cloud to enhance the UI of mobile devices.

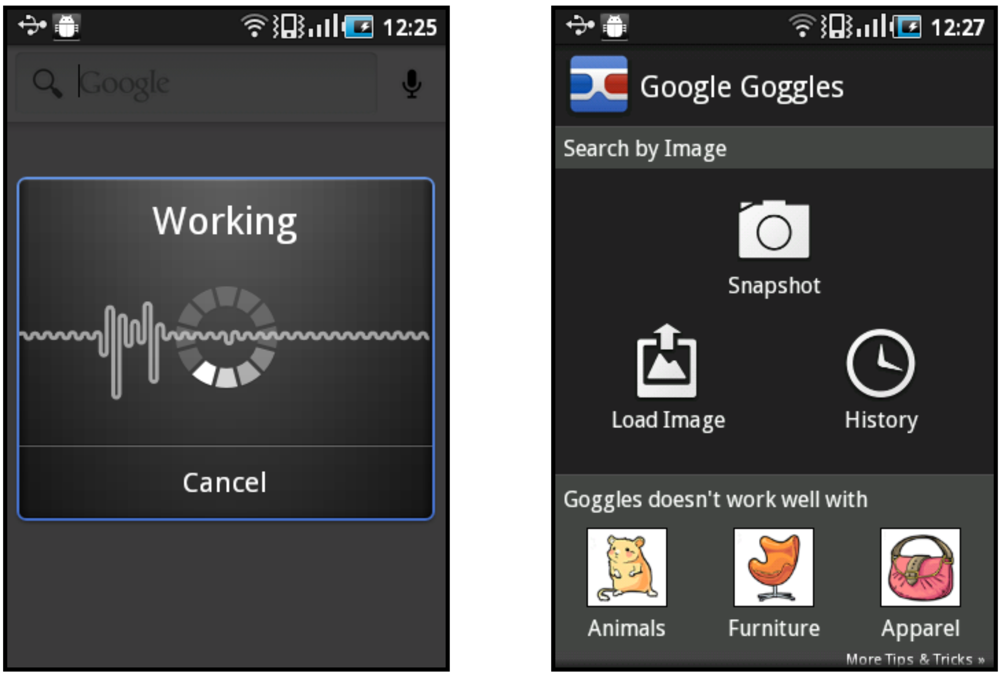

Figure 4.4: Screenshots of advanced human-computer interaction systems. On the left, the *"Google Voice Search"* app uses cloud resources to perform speech recognition on spoken search queries. On the right, the *"Google Goggles"* app allows users to search using images captured from the device's camera. Due to their high computational requirements, both the speech recognition and image processing operations are performed in the cloud.

An example of a speech recognition system enhanced through the use of cloud resources has been presented by Kondratova and Goldfarb [83]. They have suggested that, in certain applications such as mLearning, speech is a more natural method of interaction with a mobile device due to the device's small screen size [83]. One of the architectural models which they have proposed uses a remote voice server to perform the speech processing rather than performing this computation directly on the mobile device [83].

A commercial example of this type of mobile HCI technique is the *"Siri"* voice interaction system integrated into the Apple iPhone 4S device [84]. This system uses a combination of natural language processing and remote computation to provide a flexible voice interface for the mobile device [84]. Due to its high computational requirements, this type of system would not have been feasible without the augmented capacity provided by the cloud resource. In a similar manner, the processing required for various other forms of advanced mobile HCI could be performed by the cloud resource.

Cloud resources can be used to enable new types of mobile HCI for specific purposes. In particular, there has been a focus on the use of mobile devices for assisting people with disabilities. Ghaziasgar and Connan have discussed the design of a mobile front-end interface for a cloud-based system which translates sign language into English [85]. Although this system is accessed through a mobile device, the required image processing is performed on remote computational infrastructure such as a cloud resource [85]. Another example of this type of system has been designed by Angin et al. to assist the visually impaired in navigating unfamiliar environments [86]. Using the camera of a mobile device, this system identifies important features in the user's immediate surroundings [86]. The computational requirements of this near real-time image processing exceed the capacity of most MCCDs and so a cloud resource is used to perform this processing remotely [86]. This system benefits from the use of the cloud resource as a communication proxy by including information from various online mapping and route planning services in the output to the user [86].

Searching for files on a mobile device is an example of a higher-level mobile HCI application which could benefit from mobile cloud computing. Lagerspetz and Tarkoma have discussed and analysed the concept of cloud-assisted mobile search in which a cloud resource assists the mobile device to perform file indexing and search operations [45]. They have concluded that file searching on mobile devices can be significantly improved through the use of cloud-based technologies [45]. Hong, Kumar and Lu have demonstrated that the efficiency of content-based image retrieval (CBIR) can also be enhanced by offloading computation to the cloud [87].

The analysis framework is applied to the advanced mobile HCI domain as follows:

Computational Requirements: In this domain, the computational requirements of certain advanced HCI operations such as near real-time speech processing or image processing may exceed the computational capacity of the mobile device and necessitate the use of cloud resources. Although the overall computational requirements in this domain are generally not as high as in the mobile scientific computing or mHealth domains, the use of mobile cloud computing can still improve the performance of various mobile HCI applications. In some cases, this technology can also enable new methods of interaction, such as the *"Siri"* voice interaction system [84], which were previously infeasible on mobile devices.

Communication Requirements: Interaction between the user and the mobile device is highly time-sensitive. Lagar-Cavilla et al. have shown that even though sufficient bandwidth may be available, high latency can have a negative impact particularly in the critical path of user interaction [88]. Although applications in this domain are unlikely to require the transfer of large quantities of data, these strict latency requirements result in high overall communication requirements.

Mobile Network Impact: The use of this type of system in a mobile network will be highly dependent on the latency of the communication links provided by the network. Latency considerations could limit this type of application to particular RATs. Since HCI systems are applicable to nearly all mobile devices and are used frequently, this type of application has the potential to place a significant load on the mobile network. Therefore, it is important to assess the total impact of this system on the network by multiplying the communication requirements by the expected number of users.

Energy Considerations: In this domain, energy considerations are dependent on the exact type of processing being performed. In some cases, energy considerations are not applicable because the operation would not be possible without the use of the cloud resource. In other cases, it is possible to achieve significant energy savings in this application domain using mobile cloud computing. In their CBIR system, Hong, Kumar and Lu have demonstrated energy savings of up to 45% using adaptive computation offloading [87]. For the process of file indexing, Lagerspetz and Tarkoma have measured energy savings on the mobile device of more than 97% by offloading the computation to a cloud resource [45].

Information Security: Similarly to the mHealth application domain, information security is critical in this domain because of the sensitive personal information involved. The direct flow of data between the user and the mobile device is inherently a secure and trusted relationship. However, by using cloud-enhanced mobile HCI techniques, this data could be transferred via a non-private network to the cloud resource. Therefore, it is critical that the appropriate security measures are used to ensure the security of any sensitive information. However, this does not imply that additional security measures are always required. By using context to identify the nature of the information involved, the system could decide on the appropriate level of security to be enforced for each particular scenario.

System Availability: The availability of systems which enhance mobile HCI is also very important because of the potential for users to become dependent on these systems. This user dependency results from both the improved ease-of-use of these systems as well as the fact that certain applications may have been designed specifically to use these capabilities. A system failure in a cloud-based mobile HCI system could affect a number of other applications or even render the mobile device largely unusable. Therefore it is important to ensure that such a system provides sufficient redundancy to meet these high availability requirements.

Application Usability: Applications in this domain are designed to enhance the usability of other functional areas of the mobile device. However, in some cases, it is also important to consider the usability of these cloud-enhanced HCI systems. For example, in the sign language translation system, the graphical output is still limited by the small screen size of the mobile device. It is also important to assess the suitability of various forms of HCI under this aspect. For example, speech recognition is not appropriate for use in quiet environments such as libraries. By considering the various use-cases it will be possible to ensure that the system operates within all applicable usability constraints.

The advanced HCI domain encompasses all systems which facilitate interaction between the user and the mobile device. Techniques such as speech recognition (voice control), handwriting recognition and gesture commands are often more natural methods of interacting with mobile devices and in same cases also provide an improved means of interaction for people with disabilities. However, advanced HCI systems have high computational requirements in order to achieve the required level of accuracy. Due to the event-based nature of user interaction, low-latency communication is critical and thus communication requirements are also high. Since these systems could attract a large number of users, it is important to assess the overall impact on the mobile network. It has been shown that significant improvements in energy efficiency can be achieved in this domain through the use of cloud resources. Information security is a critical consideration since these systems could involve sensitive user information. System availability is also important because of potential user dependency on these systems. Although these applications are designed to enhance usability, it is also important to ensure that the applications themselves can operate within the usability constraints of a mobile device.

4.6 Entertainment and Multimedia

MCCDs are increasingly being used as mobile entertainment devices. Activities of this type generally involve some form of multimedia content including images, audio and video. This application domain encompasses all systems which focus on multimedia and entertainment activities. However, mobile gaming, which is a special case of this domain, is analysed as a separate domain in the next section. A characteristic of mobile gaming is that the user interacts with the mobile device very frequently. In contrast, the mobile entertainment and multimedia domain is characterized by a much lower degree of user interactivity. For example, mobile video streaming and mobile graphical gaming both incorporate multimedia elements but the gaming scenario will involve a much higher level of user interaction because of its event-based nature.

Mobile users can be both producers and consumers of multimedia content which is often distributed via the internet. The production of multimedia content involves the acquisition of data through a capture peripheral, such as the mobile device's video camera, as well as some form of data processing including image manipulation, audio processing or video editing. The consumption of this content on mobile devices can also involve computationally intensive operations such as rendering high resolution video. Multimedia content generally consists of a significant amount of data. Two examples of mobile entertainment and multimedia apps are shown in **Figure 4.5**. This figure shows the *YouTube* cloud-based video streaming application and the *BBC iPlayer* app which provides access to television and radio content via a mobile device.

Mobile video is already the largest consumer of data in mobile wireless networks and is predicted to experience significant growth in the coming years [89]. The *"Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2010 - 2015"* predicts that in 2015 mobile video will account for 66.4% of all mobile network traffic and will exceed 4 exabytes per month as shown in **Figure 4.6** [89].

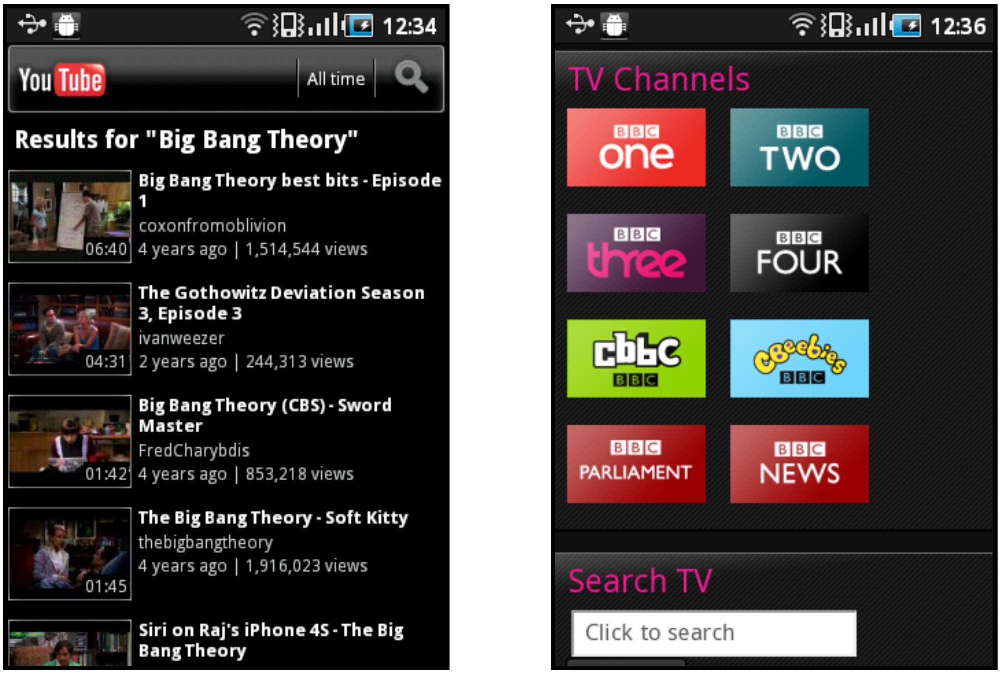

Figure 4.5: Screenshots of mobile entertainment and multimedia apps. On the left, the *"YouTube"* mobile app allows users to stream video from the cloud to their mobile devices. On the right, the *"BBC iPlayer"* app provides access to streaming television and radio content on mobile devices.

This domain also includes applications used to browse multimedia-rich web sites. With the emergence of new web technologies such as HTML5 which are designed for multimedia content, it is likely that even more web sites will provide a multimedia-rich user experience. However, this increase in multimedia content will also increase the computational requirements of client applications such as mobile web browsers.

Shen et al. have demonstrated a mobile web browser application which uses the cloud as a communication proxy [90]. The objective of this system is to offload the processing of web pages (including multimedia elements) from the mobile device to the cloud resource [90]. Experimental testing has shown that this approach reduces the time taken to load web pages on the mobile device [90]. A recent commercial example of this type of application is the *"Amazon Silk"* cloud-accelerated web browser provided on the Amazon Kindle Fire

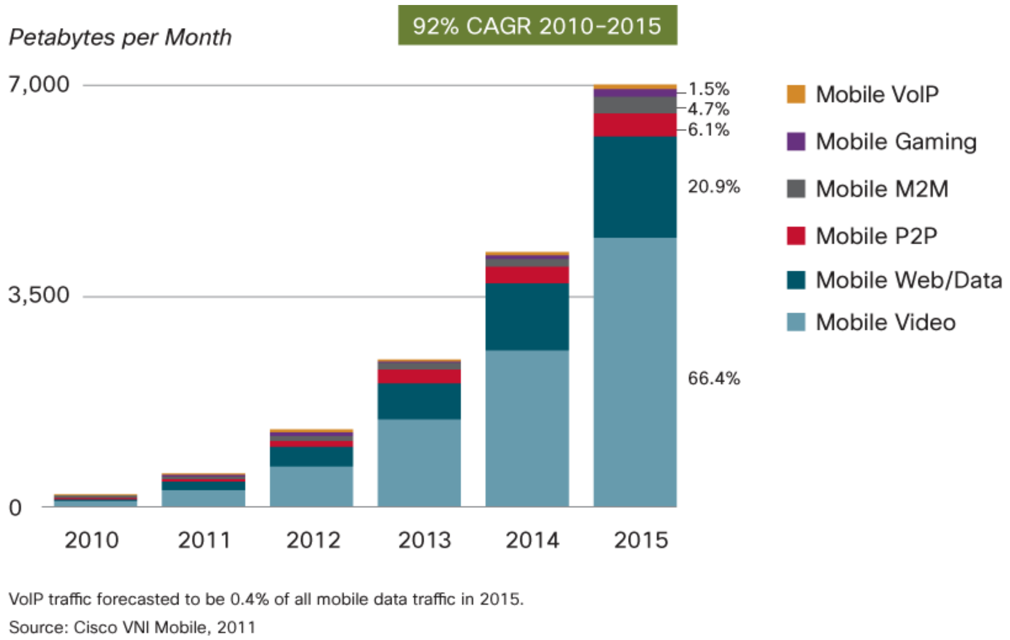

Figure 4.6: Predicted monthly data traffic per application carried by mobile wireless networks. Mobile video is already the largest consumer of mobile data and is predicted to increase significantly in the coming years.
(Source: Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2010 - 2015 [89].)

mobile device [91]. This system is described as a *'split-browser'* and includes components on both the mobile device and its associated cloud resource [91]. Again the cloud is used as a communication proxy but in this case the computation can be dynamically distributed between the mobile device and the cloud [91]. These cloud-enhanced mobile web browsers are examples of the application-specific approach to mobile cloud computing as explained in **Section 5.4**.

Application of the analysis framework to the entertainment and multimedia domain highlights the following points:

Computational Requirements: Although the computational operations required to process multimedia content may have a relatively low complexity, the quantity of data to be processed generally results in high computational requirements. When processing multimedia content, there is a direct and often very apparent relationship between the computational capacity of the device and the output quality. Although a mobile device may provide sufficient computational capacity to process this content, the use of cloud resources to augment this capacity could result in increased output quality and computational efficiency.

Communication Requirements: In this domain, communication requirements are dominated by bandwidth rather than latency considerations because of the large quantities of data involved. Due to the low degree of user interactivity, data can be buffered on the mobile device to compensate for communication latency and improve output quality. In this aspect, the use of the cloud resource as a communication proxy can significantly increase the communication efficiency of the system. For example, when a video is streamed from a network end-point to the mobile device, it is possible that the resolution of the video will exceed that of the mobile device display resulting in the transfer of unnecessary data over the wireless communication link. When used as a communication proxy, the cloud resource could obtain the original content and reformat it specifically for the target mobile device thus minimizing communication requirements. Jin and Kwok have suggested a similar idea and have also proposed that this model could reduce energy consumption, augment the computational capacity and improve the communication efficiency of mobile devices [92].

Mobile Network Impact: The impact of mobility on multimedia applications would most likely be manifest in temporary unforeseen disconnections from the network. However, the buffering of content eliminates the effects of these disconnections and makes this application domain suitable for use in mobile networks. This domain could potentially have a very high impact on the mobile wireless network because of the large quantity of data involved and the large number of potential users. Currently, mobile video applications generate more mobile network traffic than any other type of application [89]. Since the use of cloud resources will reduce the amount of data transferred over the wireless communication network, mobile cloud computing will lessen the impact of this application domain on mobile networks.

Energy Considerations: According to the energy cost model proposed by Kumar and Lu [1], the high quantity of communication involved in this domain could limit any potential increase in energy efficiency. Although on-device multimedia processing requires a significant amount of energy, an equal or greater amount of energy could be required to transfer the unprocessed content to the cloud in certain circumstances. Therefore the offloading decision depends on variable factors such as the available network bandwidth. However, if the multimedia content originates from an external network end-point, the use of the cloud resource as a communication proxy can provide significant improvements in energy efficiency. Using this approach, data obtained from an external network end-point is processed in the cloud before being transferred to the mobile device. If the cloud resource is using energy efficient infrastructure (as is the case in most public clouds), this approach could also reduce the total energy consumed by the whole system. Therefore, depending on the source and destination of multimedia content, mobile cloud computing has the potential to improve energy efficiency in this domain.

Information Security: In this domain, the importance of the information security aspect depends on the type of content being transferred or processed. A significant amount of the multimedia content consumed on mobile devices originates from network end-points in the form of video streaming. The implications of a security breach in such applications are no more serious than a general security breach on any personal system. Through such a breach, a user's privacy would be compromised but no sensitive information would be revealed. In this case, the mobile cloud computing approach does not require any additional information security measures above those already in use for the protection of user privacy.

System Availability: The availability requirements of applications in this domain are not critical. Although the use of mobile cloud computing enhances the quality and improves the efficiency of such applications, it is not required for the system to function correctly. The effects of network variability in current mobile multimedia applications will remain largely unchanged using mobile cloud computing. For example, when streaming video over the network, poor network connectivity may necessitate the use of a lower resolution in order to maintain an acceptable frame-rate. In a system based on mobile cloud computing, these same constraints are applicable.

Application Usability: In this domain, application usability is an important consideration because it influences both the quality of the output and the efficiency of the application. For example, in a mobile video application, hardware characteristics such as screen size and resolution limit the maximum appreciable output quality. Although higher quality content may be available, it would not be beneficial to stream this to the mobile device. By using the cloud resource to convert the content to match the capabilities of the mobile device, the communication efficiency of the system can be improved. Conversely, low-quality content may need to be up-scaled in the cloud before being displayed on devices with larger screens such as tablet PCs to ensure that an acceptable level of output quality is maintained.

In the multimedia and entertainment application domain, the primary focus is on the content used by the applications rather than the applications themselves. Although the computational operations involved in processing this content are not very complex, the volume of content results in relatively high computational requirements. This large quantity of data also leads to high communication requirements which are dominated by bandwidth rather than latency characteristics because of the low degree of interactivity. The mobile network impact of this domain could potentially be very high due to the broad target audience. Using the energy cost model, offloading decisions based on energy considerations will depend on the exact quantity of data to be transferred and the prevailing network characteristics. However, the use of the cloud resource as a communication proxy could reduce energy consumption, improve the communication efficiency and reduce the network impact of these systems. Information security and system availability considerations are not critical in this domain due to the nature of the service being provided. Application usability must be taken into account because device hardware characteristics may limit the quality of the output and necessitate preprocessing of content in the cloud.

4.7 Mobile Gaming

The mobile gaming application domain encompasses all gaming-related activities on mobile devices. The range of available gaming applications is very broad and so the generality of conclusions in this domain is inherently limited. However, for the purposes of this analysis, gaming applications can be characterized as multimedia-intensive applications which exhibit a high level of user interactivity. Although general multimedia applications are analysed in the previous section, the interactivity criterion in the mobile gaming domain modifies the outcome of the analysis. The event-driven nature of gaming essentially eliminates the possibility of buffering multimedia output and so adds additional performance requirements. Apart from the multimedia processing, many games also include complex game-related algorithms (game logic) to enhance the gaming experience.

Barboza et al. have proposed an architecture for on-demand gaming on devices with limited computational capacity (including mobile devices) [93]. In this architecture, the majority of computation is performed using cloud resources [93]. Prototype testing has shown that cloud computing infrastructure can be successfully used in this application domain [93].

The *"Games@Large"* research project is an example of a distributed computation system which aims to bring PC quality gaming to resource-constrained devices [94]. Nave et al. have described how this system can stream either the game state or a fully rendered visual output depending on the capabilities of the end device [94]. There are also various commercial cloud-based gaming systems such as *"OnLive"* [95] and *"Gaikai"* [96]. Using a distributed execution approach, these services provide users with a means to play certain PC games on a variety of consumer devices including smartphones and tablet PCs [95][96]. Both of these services are examples of the successful implementation of cloud-based mobile gaming. **Figure 4.7** shows a screenshot of a multimedia-intensive game being played on a mobile device using the *"OnLive"* system.

Figure 4.7: Screenshot of multimedia-intensive game being played on a mobile device powered by the *"OnLive"* cloud-based gaming system [95].

Luo has investigated the use of the cloud computing paradigm in the context of virtual reality or augmented reality applications [97]. These applications have fundamentally similar requirements to mobile games including high graphical throughput, low latency multimedia rendering and frequent user interaction [97]. Luo has concluded that although this is a relatively new approach, the cloud computing paradigm will be beneficial to these types of applications in the near future [97].

Wang and Dey have presented a technique to address the constraints of communication bandwidth in mobile gaming applications [98]. This approach involves dynamic adaptation of the rendering of graphical output based on network conditions [98]. Using a commercial mobile wireless network based on Universal Mobile Telephony System (UMTS) technology, it has been demonstrated that this strategy can significantly improve usability compared to a static approach [98]. This leads to the concept of application partitioning which is discussed in **Section 5.3**.

The use of the analysis framework in the mobile gaming domain gives the following results:

Computational Requirements: This domain is characterized by very high computational requirements due to the high quantity of multimedia content as well as the various game-related algorithms involved. The output quality of the application is proportional to its computational requirements. Greater computational capacity allows the rendering of richer multimedia content and the use of more sophisticated game-related algorithms. Furthermore, the event driven nature of gaming means that computation must be performed in near real-time. Although the current generation of mobile games provides an attractive user experience, this could be significantly enhanced using the augmented computational capacity provided by the cloud.

Communication Requirements: The quantity of data exchanged between the mobile device and the cloud resource depends on the nature of the application. The generalization of games as multimedia-intensive applications with a high level of user interactivity leads to very high communication requirements. Jarschel et al. have explained that unlike conventional video streaming or web applications, cloud-based gaming requires both a relatively high downlink bandwidth and low communication latency [99]. In this domain, almost all communication takes place directly between the mobile device and the cloud resource and therefore there is no benefit in using the cloud resource as a communication proxy. Communication latency also has a direct effect on the output quality of the application. In order to reduce this communication latency, it may be desirable to situate the cloud resource in close geographical proximity to the mobile device as proposed by Satyanarayanan [30].

Mobile Network Impact: Due to the high communication requirements of this domain, mobile gaming applications could be limited to certain types of RATs in order to ensure the required bandwidth and latency characteristics. Similarly to the multimedia domain, the broad appeal of mobile gaming could result in a very large user base. This large number of users combined with the high quantity of multimedia data transferred could place a significant load on the mobile network. However, the system also inherently exhibits some degree of self-regulation because of the strict communication latency requirements. If the network were to become congested, communication latency would increase, making it unsuitable for gaming. The applications in this domain would then automatically reduce their usage thus returning the network to it previous state. Apart from restricting these applications to specific RATs, it may also be necessary to categorize and prioritize different classes of network traffic during peak periods.

Energy Considerations: Due to the computationally intensive nature of mobile gaming, there is significant potential for energy saving through the use of cloud resources. Using the *"MAUI"* system, Cuervo et al. have shown that remote execution reduced the energy consumption of a graphically-intensive video game by 27% and a mobile chess game by 45% [31]. There is a difference in energy saving because the chess game involves less multimedia content but more algorithmic logic than the video game and can therefore offload more computation using less communication. Again, the energy cost model can be used to make offloading decisions in this type of scenario. Given the correct conditions, there is significant potential to reduce energy consumption in this application domain.

Information Security: In this domain, it is not likely that applications will involve highly sensitive information. Therefore, no additional security measures are required beyond those already used for securing network communication. It may be argued that some mobile games involve online payment for the game itself or for specific features. However, this transfer of payment details over the network is not a new concept and existing security measures have been designed to address this scenario. The advent of cloud-based mobile gaming does not increase risk or introduce any new vulnerabilities from a security perspective. Jarschel et al. have pointed out that cloud-based games are advantageous to game developers because the cloud can be used to reduce software piracy or other unauthorized use of the software [99].

System Availability: In comparison to other application domains analysed in this chapter, mobile gaming applications generally do not have high availability requirements due to the nature of the service. However, since this is also one of the most highly commercialized domains, periods of system unavailability would have a significant impact on revenue for the game's provider. Although the mobile cloud computing paradigm introduces networking and cloud elements into the system, it may still be possible for games to continue functioning if these resources become unavailable. In this reduced functionality mode, a game could temporarily decrease output quality or use simplified game logic algorithms in order to reduce its computational requirements. This could be used as a fall-back mode for short periods when the mobile device temporarily loses connectivity to the cloud resource.

Application Usability: The design of gaming applications is constrained by the limited UI capabilities (input and output) of mobile devices. Due to the high quantity of multimedia content involved, factors such as screen size and resolution impose a maximum limit on the quality of the output of mobile gaming applications. Although the use of cloud resources makes it possible to remotely generate and render highly detailed graphical output, this level of detail may not be visible on a mobile device's screen. In some cases, attempting to display too much information may decrease the overall user experience. In a cloud-based gaming application, the network and cloud elements could also affect usability by introducing perceptible delays in the game. Jarschel et al. have explained that, due to the differences between cloud-based gaming and traditional gaming, a new Quality of Experience (QoE) metric is required [99]. They have proposed a QoE metric and demonstrated the use of this through a user survey [99]. Based on the user experience data obtained, they have concluded that cloud-based gaming is a viable option for the future [99].

Mobile games are usually multimedia-intensive applications characterized by a high degree of user interactivity. The use of cloud-based resources to bring gaming to resource-constrained devices has been explored in recent scientific literature and forms the basis of multiple commercial systems. The computational requirements of this domain are very high due to the high quantity of near real-time multimedia processing required. Output quality is often directly related to the amount of computation required. Communication requirements are high because games require both high bandwidth and low latency in the connection between the mobile device and the cloud resource.

Due to their broad user base, mobile gaming applications could place a significant load on the mobile network. However, the inherent self-regulation of these systems would prevent network congestion. It has been shown that significant improvements in energy efficiency are possible in this domain depending on the nature of the game. In this domain, information security is not critical since the addition of the cloud element does not introduce any new security vulnerabilities. Similarly, system availability is not a primary concern as it is often possible to include a fall-back mode for temporary disconnected operation. Application usability is an important consideration because it limits the maximum appreciable output quality of the game and hence the computational requirements. Since usability could also be affected by the network and cloud elements, it is important to analyse applications using an appropriate QoE metric. Overall, this domain has significant potential to benefit from the mobile cloud computing paradigm.

4.8 Generic Analysis Process

The preceding sections of this chapter have demonstrated the process of applying the theoretical analysis framework defined in **Chapter 3** to specific application domains. One of the stated aims of the framework is that it should be sufficiently flexible so that it can be used in the analysis of different types of systems based on mobile cloud computing. The preceding analyses of different domains serve to validate this requirement.

This section expands on the preceding sections and presents a generic analysis process which explains how the framework can be used to analyse other applications or systems based on mobile cloud computing. The exact process of using this framework differs depending on the system being analysed. However, in all cases, there are certain overarching stages of the process which are present in all analyses.

Although the stages of this process form a logical sequence, the analysis often includes iterations in which the output of a particular stage leads to refinements in one of the earlier stages. This iterative approach resembles a type of agile methodology as defined in the field of software engineering [100].

The major stages of the generic analysis process for mobile cloud computing systems are:

System Identification: The first stage in the process is the identification of the system or group of systems to be analysed. This is usually based on a functional definition of the system which describes the core features or capabilities. Using this functional definition, the specific subsystems and technologies required to provide this functionality can be identified. It is important to ensure that the complete system is identified in this stage. This analysis can also be performed on groups of systems as demonstrated by the use of application domains rather than specific systems in the preceding sections of this chapter. In this case, a functional definition common to all the systems in the group must be established. In the preceding sections, the functional definition of each application domain is presented as part of the introductory discussion in each section.

System Verification: The second stage in the process is to verify that the system being analysed is indeed based on the mobile cloud computing paradigm. Mobile cloud computing must be distinguished from a similar paradigm in which computation is distributed amongst a group of mobile devices in close physical proximity. As explained in **Section 5.2**, this type of system architecture is known as mobile grid computing. Although they achieve similar objectives, systems based on the mobile grid computing paradigm are beyond the scope of this analysis framework due to fundamental differences in system architecture. In the preceding sections, the system verification stage is documented implicitly in the initial discussion of each application domain.

Mobile-Cloud Relationship Analysis: The third stage of the process consists of an analysis of the relationship between the mobile device and the cloud resource. This aims to determine which functional components are executed on the mobile device and which are offloaded to the cloud. In some systems, computation may be dynamically partitioned between the mobile device and the cloud as explained in **Section 5.3**. For these systems, the analysis must include both possibilities for the relevant components. In this stage, the flow of information between the mobile device and the cloud resource is also determined. In the preceding sections, the results of this third stage are explained in the discussion of each application domain.

Analysis of Aspects: The core analysis of the system based on the aspects defined in the framework is the fourth stage of the process. As explained in **Chapter 3**, each aspect of the framework consists of a set of major considerations. In this stage, each of these considerations is evaluated in the context of the system under analysis. The analysis framework has been designed primarily to facilitate qualitative analysis of systems or groups of systems. If a functional prototype of the system is available, it is possible to supplement the main analysis with additional quantitative benchmarks for certain aspects. However, as demonstrated by the preceding sections of this chapter, qualitative analysis is sufficient to facilitate the desired comparisons between different systems. The results of this core stage of the analysis are documented specifically for each aspect.

Analysis of Interdependencies: For each aspect of the framework, a number of interdependencies with other aspects have been identified. As explained in **Section 3.9**, it is critical to include these interdependencies in any analysis of mobile cloud computing systems. In this stage, the results of the analysis for each aspect are evaluated with reference to the relevant interdependencies. If required, outcomes of the previous stage are refined based on these relationships between aspects. The results of this stage of the process are documented with the results for each aspect from the preceding stage.

Conclusion of Analysis: The final stage of the analysis process is to identify the most significant insight and conclusions for a particular system or group of systems. These arise from the analysis of the specific aspects and interdependencies in the preceding two stages. In the analyses of the various application domains in this chapter, the conclusions for each domain are documented and discussed at the end of each section.

Using the process presented above, the framework defined in **Chapter 3** can be used in the analysis of other systems based on the mobile cloud computing paradigm.

4.9 Conclusion

The analysis described in this chapter fulfils the third requirement of this research as specified in **Chapter 1**. It demonstrates the use of the theoretical analysis framework, described in **Chapter 3**, in analysing various mobile application domains.

As stated in **Section 4.1**, the purpose of this chapter is to demonstrate the analysis of various mobile application domains using the theoretical framework. These analyses serve as validation of the framework itself. They also provide useful insight and conclusions relevant to each application domain as summarized in this section. This chapter demonstrates the process of using the analysis framework and, in particular, the generic analysis process described in **Section 4.8** explains how the framework can be used in the analysis of other mobile cloud computing systems.

In this chapter, individual mobile apps are grouped into application domains based on their functional characteristics. This allows the insight and conclusions obtained to be generalized to all apps in a particular domain and avoids ambiguity in the analysis of multifunctional apps. The application domains featured in this analysis have been selected based on the degree of benefit they derive from the use of mobile cloud computing and have been sourced from recent scientific literature as well as commercial endeavours. However, this is not an exhaustive list of application domains since new functionality is constantly being developed.

Some of these application domains have been discussed in recent scientific literature in the context of mobile cloud computing. However, these are usually analysed using only a limited subset of criteria. The use of the analysis framework provides a comprehensive qualitative analysis, supported by literature and examples, showing how each aspect of an application domain is affected by mobile cloud computing. This highlights areas which benefit from the use of this technology but also identifies potential issues which must be adequately addressed in the design of the application. The use of this common analysis framework facilitates comparisons between different application domains based on specific aspects of the framework.

In summary, for each aspect of the analysis framework, the following important conclusions have been identified from the analyses described in this chapter:

Computational Requirements: The computational requirements of an application domain are one of the most important factors in determining the degree of benefit the domain derives from the use of mobile cloud computing. In some cases, these computational requirements exceed the computational capacity of the mobile device thus necessitating the use of cloud resources. This occurs in the mobile scientific computing and mHealth domains due to the complexity and quantity of computational operations required and the quantity of data to be processed. Even if the mobile device has sufficient computational capacity, the computational requirements of these application domains are very high and so the use of the cloud resource can significantly increase the computational performance of the system. The computational requirements of the mLearning domain are the lowest out of this comparison since they primarily involve reformatting and storing educational content. In the advanced HCI domain, the complexity of the computations means that cloud resources can be used to improve performance or provide new functionality. The computational requirements of the multimedia domain are relatively high due to the large quantity of multimedia processing involved and are even higher in the mobile gaming domain which adds game-related algorithms and requires near real-time computation. In both of these domains, the output quality is directly proportional to the computational requirements. Therefore, the use of the cloud resource to augment the computational capacity of the mobile device leads to an improved level of output quality and a better user experience.

Communication Requirements: The communication requirements vary between application domains but are also important in determining the degree of benefit provided by mobile cloud computing. In the mobile scientific computing and mHealth application domains, the communication requirements are relatively low because, although the quantity of data may vary, there are no strict time constraints on the transfer of this data. In the mHealth domain, the cloud resource can also be used to facilitate the transfer of information to a medical professional. Communication requirements are important in the mLearning domain due to its focus on content and collaboration. In this domain and in the multimedia domain, the use of the cloud resource as a communication proxy yields significant increases in communication efficiency. Bandwidth and latency requirements strongly influence the overall communication requirements in the HCI, multimedia and mobile gaming domains. In the HCI domain, the latency of the communication is critical even though the quantity of data transferred is relatively low. Conversely, the multimedia domain requires high bandwidth to transfer large quantities of multimedia content but can mitigate the effects of communication latency through content buffering. The mobile gaming domain has the highest communication requirements because it requires both high bandwidth and low latency communication between the mobile device and the cloud resource. In order to reduce communication latency, the use of distributed *cloudlets* in close geographical proximity to the mobile device has been suggested by Satyanarayanan [30].

Mobile Network Impact: This aspect covers both the impact of the mobile network on each application domain as well as the impact of each domain on the network. The mobile scientific computing and mHealth domains do not have a significant impact on the mobile network because they are infrequently used and can utilize low-bandwidth data transfer channels. In these two domains, the possibility of disconnection from the network during a long-running process must be adequately addressed. For applications described in the mLearning and multimedia domains, the use of cloud resources can improve communication efficiency and decrease the load on the network. This is particularly important in the multimedia domain which is already the largest source of mobile network traffic [89]. Since the mLearning and multimedia domains are largely focused on content, it is possible to mitigate mobile network effects such as communication latency through content buffering. The HCI and mobile gaming domains would likely be restricted to certain RATs because of their requirements for low communication latency. The total communication requirements of all users must be considered in the advanced HCI, multimedia and gaming application domains to ensure that sufficient network capacity is available. Mobile gaming applications cannot use the network if it is in a congested state and therefore exhibit a degree of inherent self-regulation which assists in limiting the network impact of these applications.

Energy Considerations: Improvements in energy efficiency are one of the most often cited benefits of mobile cloud computing. The importance of energy efficiency is a direct consequence of the finite energy storage capacity of mobile devices which is one of the fundamental constraints of mobility [14][17]. Energy considerations are used to make computation offloading decisions and to quantify the degree of benefit provided by mobile cloud computing. An energy cost model (similar to that discussed by Kumar and Lu [1]) can be used to analyse energy considerations in distributed computing systems. The majority of applications in the mobile scientific computing domain would not be feasible without the use of cloud resources and thus energy is not a primary consideration in this domain. The same is true for certain applications within the mHealth domain although for others there is significant potential to improve energy efficiency by distributing computation between the device and the cloud. In the advanced HCI domain, energy savings depend on the type of computation involved and significant improvements can be achieved in the more complex higher-level HCI operations. The best candidates for improvements in energy efficiency are the mLearning, multimedia and mobile gaming domains. In these domains, various complex or high-quantity computational operations can be performed in the cloud. Although this requires the mobile device to expend energy on communication, the reduction in computation leads to overall energy savings. In these application domains, the energy cost model shows how the use of mobile cloud computing can significantly increase the energy efficiency of mobile devices and lead to system-wide energy savings.

Information Security: In terms of information security and privacy considerations, application domains vary depending on the type of information involved. The mobile scientific computing and mLearning domains most often utilize open-access or pubic domain content such as public data-sets or open educational resources. The combination of this type of information and the collaborative nature of these domains means that information security is not a primary consideration. This does not mean that these systems should not have certain security measures but rather that the use of existing device and network security protocols is sufficient in these domains. In contrast, information security is a critical requirement in the mHealth and advanced HCI domains because these involve highly sensitive personal information. When medical information is stored or processed in electronic form, regulatory statutes enforce certain levels of security. In these two domains, advanced security services must be an integral part of an application's design. The multimedia and mobile gaming domains fall between these two extremes because they do not involve sensitive information but still require the protection of user privacy. This can be achieved using various existing security services and should be considered in the design of the applications. Research has shown that it is possible to address the information security considerations of mobile cloud computing using the appropriate security technologies. An example of such a solution is the security management model proposed by Nkosi and Mekuria [73].

System Availability: In the above analysis, the importance of system availability varies depending on the application domain. The aim is to minimize system unavailability in all domains but the consequences of such unavailability are different for each domain. Due to the nature of the mobile scientific computing domain, temporary system unavailability would not have serious consequences. By comparison, in the mHealth and advanced HCI domains, it is possible that users may become completely dependent on these applications and so unavailability could result in serious medical consequences or the mobile device becoming unusable. Applications in these domains therefore require a sufficient level of redundancy to mitigate against this risk. The actual level of availability would probably be specified through some form of service level agreement. In the mLearning domain, the success of the system is influenced by the level of system availability. However, since this is not a highly commercialized domain, the risk and consequences of system availability must be balanced with the increased resource costs required to achieve higher availability. On the other hand, the multimedia and gaming domains are highly commercial in nature and so system unavailability has a negative financial impact for the service providers. It is possible to achieve higher levels of system availability by using redundant subsystems or components with specified availability guarantees. However, the selection of these subsystems and components is often a financial decision which is beyond the scope of this analysis.

Application Usability: In the application domains analysed in this chapter, usability considerations are important because they impose constraints on the transfer of information between the user and the mobile device. However, the exact nature of these constraints varies between the different domains. In the mobile scientific computing domain, UI limitations such as small screen sizes limit the quantity of information that can be output to the user. The small physical device size also limits the input of information but this limitation can be eliminated by obtaining the input data from sensor peripherals connected to the device. In the mHealth domain specialized physiological sensors are used to obtain input data and the output information is likely to be presented in a highly condensed form. Although applications in the HCI domain are designed to enhance usability, these applications themselves must also be capable of functioning using only the limited UI hardware capabilities of mobile devices. In the mLearning, multimedia and mobile gaming domains, usability considerations mainly limit the output of information to the user in terms of quality. Although it may be possible to source or generate higher quality content using cloud resources, this would not improve output quality. Increasing output quality generally leads to higher computational and communication requirements and therefore increases energy consumption. By taking usability considerations into account, it is possible to achieve a balance between these factors in order to maximize the efficiency of the system whilst operating within the UI constraints of a mobile device.

The analyses presented in this chapter demonstrate the benefits of using the theoretical analysis framework defined in **Chapter 3**. By performing a comprehensive analysis of all aspects in a particular application domain, it is possible to identify both the implications as well as the benefits of the use of mobile cloud computing in that domain. Furthermore, the use of this analysis framework facilitates qualitative comparisons between different application domains based on specific aspects of the framework. The insight and conclusions obtained through these analyses can be fed back into the design process in order to enhance future applications. The generic analysis process presented in **Section 4.8** facilitates future use of the framework in the analysis of other systems based on the mobile cloud computing paradigm.

Chapter 5

Implementation Approaches for Mobile Cloud Computing

5.1 Introduction

The previous chapter demonstrated the analysis of various mobile application domains using the theoretical analysis framework. As highlighted by some of the scenarios presented, it is also important to consider the underlying mechanisms of these systems. This chapter completes the core technical analysis of this research by analysing various approaches to the implementation of mobile cloud computing systems. This fulfils the fourth requirement specified in **Chapter 1**. The main analysis sections in this chapter are preceded by two preparatory sections which provide relevant background information and contextualize the content of this chapter.

In the first preparatory section, **Section 5.2**, the differences between mobile grid computing and mobile cloud computing are discussed. These two concepts exhibit various similarities in terms of purpose and functionality. However, due to their fundamental differences in architecture, it is more accurate to describe these as two separate sub-paradigms rather than different implementation approaches.

The second preparatory section, **Section 5.3**, explains the concept of *'partitioning'*. This relates to the division of computation between the mobile device and its associated cloud resource. This concept is relevant to all implementation approaches discussed in this chapter. Partitioning is an ongoing research field supported by an expanding base of scientific literature. The full extent of partitioning research is beyond the scope of this work but **Section 5.3** presents the background information necessary for the core analyses in this chapter.

The implementation approaches analysed in this chapter are sourced from recent scientific literature as well as commercial endeavours. The information obtained from scientific literature ranges from suggestions based on theoretical analysis through to prototype implementations with performance measurements. The commercial endeavours are currently all being operated as production-quality systems but inherently lack the technical details necessary for in-depth quantitative analysis. Therefore, in order to include a broad range of implementation approaches (both scientific and commercial), a high-level systems analysis is used in this chapter. The theoretical analysis framework is not used directly as in the previous chapter because of the lack of in-depth technical information in some aspects. However, various considerations from that framework are used in this analysis where applicable.

In a similar manner to the previous chapter, the implementation approaches in this chapter are divided into categories based on their architectural characteristics. Since there is no formal categorization system for this type of analysis in scientific literature, the categories used in this chapter have been constructed specifically for this research. This categorization is based on the fundamental architectural characteristics of mobile cloud computing systems. By analysing these categories rather than the individual approaches, it is possible to provide insight and conclusions which are applicable to a number of different systems. Therefore, the purpose of this chapter is not to give an exhaustive list of implementation approaches but rather to identify, analyse, and comment on these major categories of systems.

The three categories into which the implementation approaches are divided in this research are:

- **Application-Specific Cloud Services**
- **Multi-Application Cloud Services**
- **Multifunctional Cloud Resources**

These three categories of implementation approaches are explained and analysed in **Section 5.4**, **Section 5.5** and **Section 5.6** respectively. The conclusions and recommendations specific to each category are given at the end of the relevant section and the overall summary and conclusions of this chapter are presented in **Section 5.7**.

5.2 Mobile Grid Computing

The term *'mobile cloud computing'* is sometimes used to refer to one of two different sub-paradigms of distributed computing systems involving mobile devices. One interpretation of this term refers to a group of mobile devices which pool their computational resources using some form of short-range wireless networking. However, the more common interpretation is a collaborative relationship between a resource-constrained mobile device and a remote high-capacity non-mobile computational resource. Based on the definition of cloud computing presented in **Section 1.3.2**, the latter interpretation is a more accurate representation of this term since it involves a remote high-capacity computational resource. For purposes of clarity, the first interpretation is referred to as *'mobile grid computing'*. Although these two sub-paradigms share the common objective of augmenting the computational capacity of a mobile device, they are fundamentally different in terms of system architecture as shown in **Figure 5.1**.

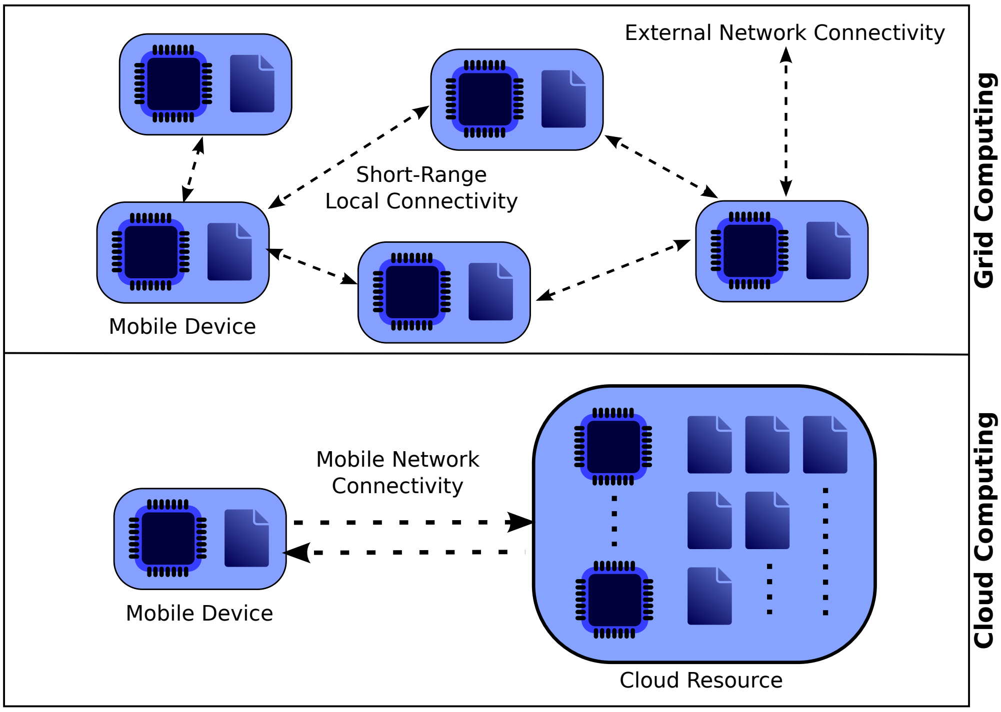

Figure 5.1: An architectural overview of the mobile grid computing sub-paradigm (top) and mobile cloud computing sub-paradigm (bottom). Mobile grid computing involves a set of similar mobile devices pooling their computational resources using short-range wireless networking. Mobile cloud computing uses the mobile communication network to link a mobile device to a remote high-capacity cloud resource.

As shown in **Figure 5.1**, mobile grid computing involves the pooling of resources between two or more mobile devices in order to obtain the computational capacity required to complete a specific task. One or more of these devices may also provide external network connectivity for the grid network but this is not mandatory. This is distinct from mobile cloud computing which pairs the mobile device with a remote high-capacity cloud resource. Although both sub-paradigms are essentially distributed computing architectures for resource-constrained mobile devices, they differ in terms of the relationship between the nodes. Mobile grid computing uses multiple resource-constrained nodes which have similar capabilities and are often located in close geographical proximity. Mobile cloud computing establishes a one-to-one relationship between a mobile device and a remote non-mobile cloud resource which has significantly greater computational capacity.

The architecture of the mobile grid computing system shown in **Figure 5.1** is very similar to that described by Masinde, Bagula and Ndegwa [69]. Their mobile grid computing system, *"MobiGrid"*, is designed to utilize low end mobile devices which are especially common in African countries. The initial applications proposed for this system include drought prediction (mobile scientific computing application domain) and preserving traditional knowledge about droughts (mLearning application domain) [69].

Huerta-Canepa and Lee have proposed a preliminary design for mobile grid computing systems (although they referred to it as cloud computing) [101]. Chu and Humphrey have demonstrated a mobile implementation of the Open Grid Services Initiative (OGSI) technology called *"Mobile OGSI.NET"* [102]. More recently, Black and Edgar have implemented a client for the Berkeley Open Infrastructure for Network Computing (BOINC) framework on a mobile device [103]. Although they have highlighted a number of challenges in the mobile grid computing sub-paradigm, their experimental results have demonstrated the feasibility of running grid systems on mobile devices [103].

One of the advantages of mobile grid computing over mobile cloud computing is that the grid architecture is self-contained and does not require a connection to an external network as shown in **Figure 5.1**. It could be argued that this is important in areas which are not covered by wireless networks. However, the devices used in these mobile grids are usually low end mobile phones since these are the most prevalent (as suggested by Masinde, Bagula and Ndegwa [69]). With advances in mobile wireless networks, it is becoming increasingly unlikely that devices such as mobile phones would be available in areas without sufficient network coverage. Therefore, although the self-contained nature of the grid is a technical advantage, it is unlikely to be manifest in real-world situations due to the nature of the mobile devices used in the mobile grid. Another perspective is that the use of the external network in mobile cloud computing carries an associated financial cost.

However, the energy consumed by the mobile device also leads to financial costs in the mobile grid computing sub-paradigm especially in areas where electricity supply is scarce. Therefore, both sub-paradigms have some form of associated financial cost but the comparison of these costs is beyond the scope of this research.

In the context of mobile healthcare (mHealth), Hoang and Chen have discussed some of the weaknesses of mobile grid computing compared to mobile cloud computing [74]. Firstly, due to the complexity of the grid architecture, it is difficult to design and deploy applications to make use of mobile grid computing [74]. Secondly, since mobile devices are already resource-constrained, it is difficult to obtain the required computational capacity in a grid consisting of only these devices [74]. As explained in **Section 2.4**, the computational infrastructure used by the cloud resource is designed to maximize energy efficiency. Therefore, it is likely that the overall energy efficiency of a mobile grid computing system will be significantly less than that of a mobile cloud computing system.

Rosado et al. have identified that security, which is already an important consideration in the general grid computing paradigm, is also very important in the design of mobile grid computing systems [104]. They have developed a methodology for the design of secure mobile grid computing systems [104].

One of the fundamental challenges of mobile grid computing is the unpredictability of the nodes participating in the grid. Park et al. have explained that the use of mobile devices in a grid computing architecture could lead to failures caused by unstable wireless connections, limited device energy capacity, low communication bandwidth and frequent location changes [105]. Since these events are largely unpredictable, additional fault tolerance is required in the system [105]. Park et al. have presented a monitoring service for fault detection in systems using this type of architecture [105].

Overall, mobile grid computing technology is seen to have potential for future development and is a topic of ongoing research. However, the fundamental differences in system architecture between mobile grid computing and mobile cloud computing lead to a number of significant differences in terms of the applications and implementation approaches of these two sub-paradigms. Therefore, although some of the relevant literature is incorporated into this research, the mobile grid computing sub-paradigm is beyond the scope of this work.

5.3 Static and Dynamic Partitioning

In a mobile cloud computing system, some computational operations are performed on the mobile device whilst others are offloaded to the cloud resource. The process of dividing computation between the device and the cloud is known as *'partitioning'*. This is an important concept in mobile cloud computing since all implementation approaches inherently use some form of partitioning. Therefore, this section aims to provide the background information necessary to contextualize the analyses presented in this chapter. Since partitioning is an ongoing research endeavour, a full review of this concept is beyond the scope of this work.

As explained by Chun and Maniatis, there are two main types of partitioning, namely, *'static partitioning'* and *'dynamic partitioning'* [52]. Between these two extremes, there are also various strategies which combine elements of both types.

In a static partitioning arrangement, the division of computation between the mobile device and the cloud resource is statically defined in the design of the application. This means that specific computational operations will always be performed by the cloud resource and all other operations will always take place on the mobile device. An example of static partitioning is the *"Google Goggles"* visual search system [106]. The mobile app is responsible for capturing and preprocessing an image which is then transferred to the cloud. The actual image search is always performed in the cloud and the results are returned to the mobile device.

The advantage of static partitioning is that the designers of the application specify the exact division of computation between the mobile device and cloud resource. The partitioning scheme is therefore based on in-depth domain-specific knowledge about the application. Another advantage is that static partitioning makes the system highly deterministic in terms of the division of computation. However, the disadvantage of this is that the system is very inflexible and unable to adapt to diversity. Flexibility is critical in the context of mobile computing because of factors such as device heterogeneity, workload diversity and the dynamic nature of mobile wireless connectivity. In this context, device heterogeneity refers to the diversity of computational capacities provided by different mobile devices. Chun and Maniatis have pointed out that many modern applications can already run unmodified on various classes of mobile devices including smartphones and tablet PCs as well as other resource-constrained computing devices such as netbooks and smart-TVs [52]. Applications based on static partitioning are unable to adapt to the different computational capacity of these devices. Workload diversity generally refers to the different types and sizes of workloads which could conceivably be input into an application. Factors such as the size and complexity of the input workload could influence the optimal partitioning scheme. Mobile wireless connectivity is inherently dynamic in nature due to a number of factors including geographic movement of the user, the available wireless coverage in specific areas and the congestion level of the network at specific times. Statically partitioned systems cannot adapt to these dynamic network conditions. Nevertheless, static partitioning is still the most logical approach for certain applications, including *"Google Goggles"* in which there is a clear distinction between the operations which must be executed locally and those which must be performed in the cloud.

In a dynamic partitioning arrangement, the division of computation takes place at runtime. The application is partitioned in different ways depending on the objectives specified for the system. These objectives could include maximizing computational performance and output quality or minimizing energy consumption on the mobile device. Due to the complexity of modern software systems, it is generally accepted that dynamic partitioning will be an automated process performed by the system. Gu et al. have explained that there are two key decision-making problems in systems of this nature [107]. The first is the timely triggering of the process and the second is the actual partitioning of the application [107]. The level of partitioning granularity varies between different systems (for example: class level, thread level or module level granularity) and is often influenced by the programming language in which the application was implemented.

Various optimization algorithms are used to automatically partition applications at runtime. For example, the *"Spectra"* system described by Flinn, Narayanan and Satyanarayanan uses gradient-descent heuristics to attempt to find the best solution [35]. Wang and Li have used parametric resource cost analysis to determine the partitioning based on a trade-off between the computational workload and the communication costs [38]. This resembles the energy cost model explained in **Section 3.5** as well as the first two aspects of the analysis framework defined in **Chapter 3**. This type of approach often requires some form of application profiling such as the estimation or measurement of the execution time of a particular operation.

Dynamic partitioning does not have the benefit of application-specific knowledge and so the outcome of the automatic partitioning may be less optimal than a well designed static partitioning scheme under a specific set of circumstances. In order to address this challenge, some systems combine the use of dynamic partitioning with additional information from the designers (as would be used in the static partitioning case). Using the *"Chroma"* system, Balan et al. have shown that it is possible to exploit application-specific knowledge within a dynamically partitioned system through the use of programmer-defined constructs which they call *'tactics'* [108]. Each tactic fully specifies one way of completing the required operation [108]. At runtime, the system makes dynamic partitioning decisions based on the information provided in these tactics [108]. Since the tactics are defined by the designers, they provide application-specific knowledge for use in the partitioning decision [108]. This is extended by the later work of Balan et al. in the development of *"Vivendi"*, a language which allows for concise description of the different tactics and fidelities of an application [109].

However, the use of *"Vivendi"* still requires input during the design phase and thus increases the complexity of the system. Furthermore, this additional information is not available for applications which have already been implemented. Xian, Lu and Li have proposed an advanced solution which combines static and dynamic partitioning without requiring any additional information from the designers and without the estimation of computation time before execution [110]. In their method, all operations are initially executed on the mobile device with specific time-outs based on a break-even analysis of the mobile-cloud combination [110]. If the local operation exceeds this break-even time-out, it is offloaded to the cloud resource [110]. Although this may appear to be sub-optimal, experimental results have shown energy savings of up to 17% compared to other approaches [110].

The choice between static and dynamic partitioning or some combination of the two ultimately depends on the requirements of the application. In some cases, the nature of application makes it particularly well suited to static partitioning. In other cases, the use of dynamic partitioning with the appropriate choice of granularity and partitioning algorithm makes the application more flexible. These flexible systems are better able to adapt to device heterogeneity and the dynamic nature of mobile wireless networks.

However, this is still an area of ongoing research in which new developments are likely to influence future implementations of mobile cloud computing. This section provides the background information necessary for the analysis of the various approaches to implementing mobile cloud computing systems. These implementation approaches are presented and analysed in the next three sections.

5.4 Application-Specific Cloud Services

This category consists of mobile cloud computing systems which are based on application-specific cloud services. For purposes of this research, an application-specific cloud service is defined as a cloud-based service designed to enhance one specific mobile app. Through such a service, this app can theoretically derive the full benefit of mobile cloud computing. These cloud services are usually proprietary or closed systems and are intentionally restricted such that they can only be used by a specific app.

Since the cloud service is designed for the exclusive use of one particular mobile app, the system designers have full control over the interaction between the mobile app and the cloud. Therefore, systems in this category are usually characterized by the use of static partitioning. In some cases multiple static partitioning schemes are provided and the system selects the best alternative at runtime. However, this is still considered to be static partitioning because the partitioning schemes were predefined during the design stage. Although technically possible, the use of dynamic partitioning has not been evident in this category of systems.

From recent literature, the *CloudTorrent* system developed by Kelenyi and Nurminen uses an application-specific cloud service approach [41]. In this system, the mobile app interacts directly with its cloud service [41]. Static partitioning is used since the roles of the app and cloud are clearly defined [41]. In this system, the cloud service retrieves a file from external internet end-points and then the mobile app downloads this data from the cloud service [41].

Two examples of commercial systems using this category of implementation approach are the *Google Goggles* visual search app from Google [106] and the *Wolfram Alpha for Mobile* scientific computing app from Wolfram Research [66]. Although the exact implementation details are not available, the operational characteristics of these systems strongly indicate the use of static partitioning. Recently, Amazon has released a Software Development

Kit (SDK) for the Android mobile OS which allows mobile applications to interact directly with the Amazon cloud services [111].

Systems in this category are relatively simple to implement because of the tight coupling between the mobile app and the cloud service. The use of static partitioning means that the communication protocols have been designed specifically for this system. The mobile app communicates directly with the cloud service as it would with other network end-points. Since all interaction with the cloud is handled by the mobile app itself, this type of approach does not require modification of the mobile OS.

Using this type of approach, a particular application can be enhanced in line with the analysis presented in the previous chapter. As explained in the previous chapter, this is highly beneficial for the mobile application and certain types of applications would not be possible without this augmented capacity. **Figure 5.2** shows an architectural overview of this category of implementation approach.

However, advances in mobile operating systems have made it possible to run multiple apps concurrently. Although only one app may be visible, this multitasking capability allows other apps to continue working in the background. As shown in **Figure 5.2**, the situation could arise where multiple cloud-enhanced mobile apps are running at the same time. Since the different mobile apps do not interact with each other, they can be said to operate using a type of silo mentality.

In a general multitasking system, multiple apps could request computational resources at the same time. The finite computational capacity of the mobile device is managed by the mobile OS and allocated to different apps using some form of scheduling algorithm. In a multitasking system which uses application-specific cloud services, each mobile app also attempts to communicate with its cloud service, leading to competition for the limited communication capacity. Since interactions with the cloud are controlled by individual applications, they cannot be managed efficiently by the mobile OS. Therefore, the OS is forced to use a naive approach and handle each

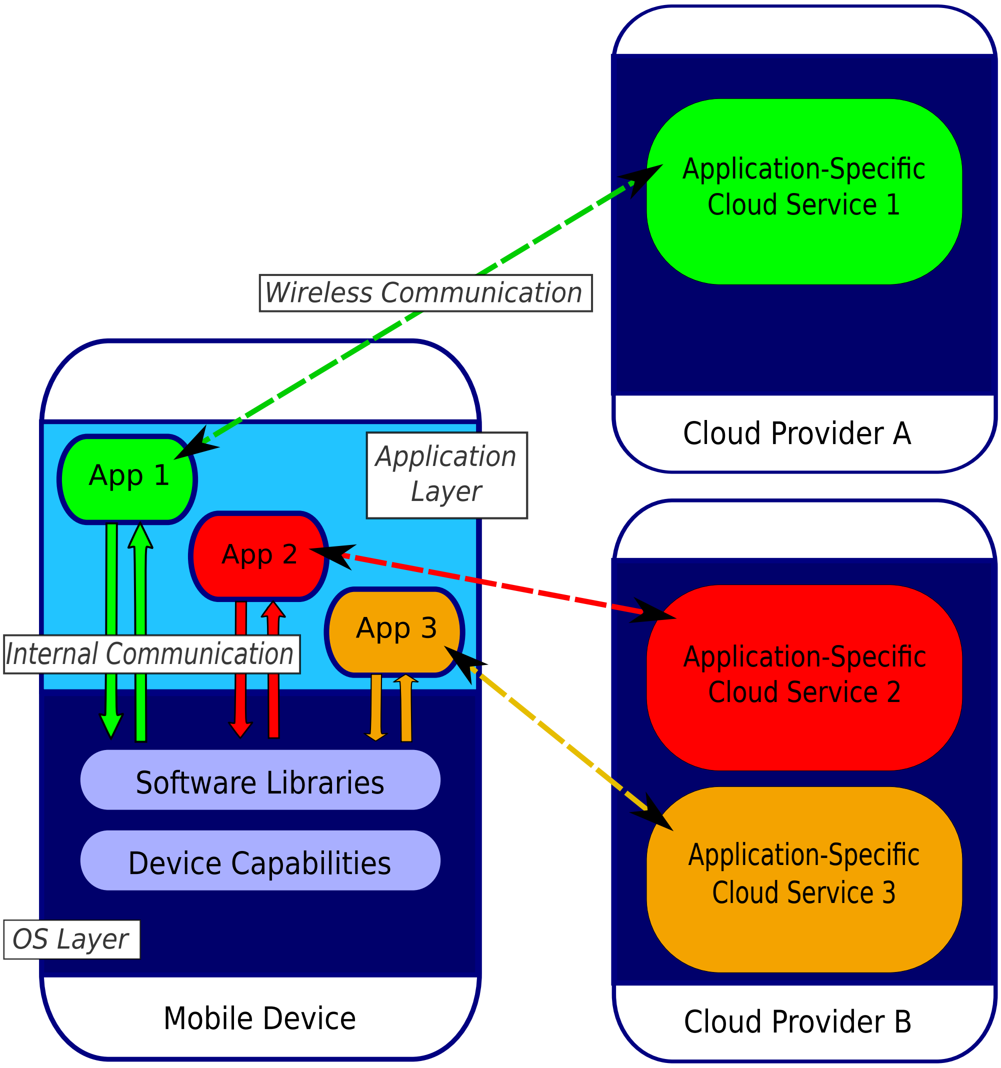

Figure 5.2: A mobile device using three separate application-specific cloud services. Each app communicates directly with a specific cloud service. Since the apps are independent of each other, multitasking results in competition for communication resources. Different cloud services may be hosted by the same cloud provider in which case multiple parallel connections will be established thus reducing the communication efficiency of systems in this category.

app-cloud interaction separately. In some cases (as shown in **Figure 5.2**), the services being accessed are provided by the same cloud provider. Due to the silo mentality of this type of approach, the mobile device would establish a separate connection to the same cloud provider for each app. These multiple parallel connections reduce the communication efficiency of the system thus placing a greater load on the mobile network and increasing the energy consumption of the mobile device.

The silo nature of this approach is advantageous from an information security perspective because each application has full control over the offloading decisions. The design of the application can use the appropriate security measures to protect any sensitive information. Also, since the applications do not interact with each other, the risk of security vulnerabilities caused by other software on the device is decreased. In terms of availability, this approach is also beneficial because the unavailability of a particular cloud service only affects a single application. However, the risks and consequences of unavailability due to failure of the communication network remain unchanged.

The proxy-based mobile web browser created by Shen et al. [90] as well as the Amazon Silk split-browser [91] could both be placed into this category because these two systems use application-specific cloud services to enhance the web browser application. However, it could also be argued that the web browser is actually a type of platform which supports other applications (web apps). In this case, the use of cloud services would benefit multiple web apps and the systems could be placed in the multi-application cloud services category discussed in the next section.

Overall, application-specific cloud services can be considered to be entry-level implementation approaches for mobile cloud computing because they are the simplest to implement. Although each service provides benefit to one particular application, it does not enhance the mobile device as a whole. This type of approach does not operate well in a multitasking environment because it leads to competition between applications for limited communication capacity and cannot be efficiently managed by the mobile OS. Although it has some advantages, the silo mentality of this type of approach makes it inefficient from a communication perspective.

5.5 Multi-Application Cloud Services

Multi-application cloud services are implementation approaches that provide specific functionality for use in multiple mobile applications. In contrast to the previous category, individual apps do not access these cloud services directly. Instead, each cloud service is accessed via its respective software component on the mobile device as shown in **Figure 5.3**. This component manages all communication with the cloud and provides specific functionality to applications via an internal software interface. The use of this management component on the mobile device means that this category is not simply an extension of the previous category.

The management component of each service interacts very closely with the mobile OS and often modifies the OS to a limited degree. For example, a service may require specific libraries to be added at the OS level so that the functionality it provides can be discovered and used by apps on the device. The extent to which modification of the OS is required depends on the type of functionality provided. This means that implementation approaches in this category are more complex than those in the application-specific category.

Multi-application cloud services use either static or dynamic partitioning depending on the type of functionality they provide. Although dynamic partitioning further increases the complexity of the system, it improves flexibility and addresses the concerns of device heterogeneity and dynamic mobile network connectivity. Since the software component on the mobile device manages the service for all applications, the benefits of dynamic partitioning are also propagated to multiple applications.

Figure 5.3 shows an architectural overview of this type of implementation approach. The software component of the service on the mobile device is shown on the border between the application layer and the OS layer. In the OS layer, this component may need to modify the OS or introduce additional components such as software libraries. It could also include an element in the application layer that interacts with users or provides status information.

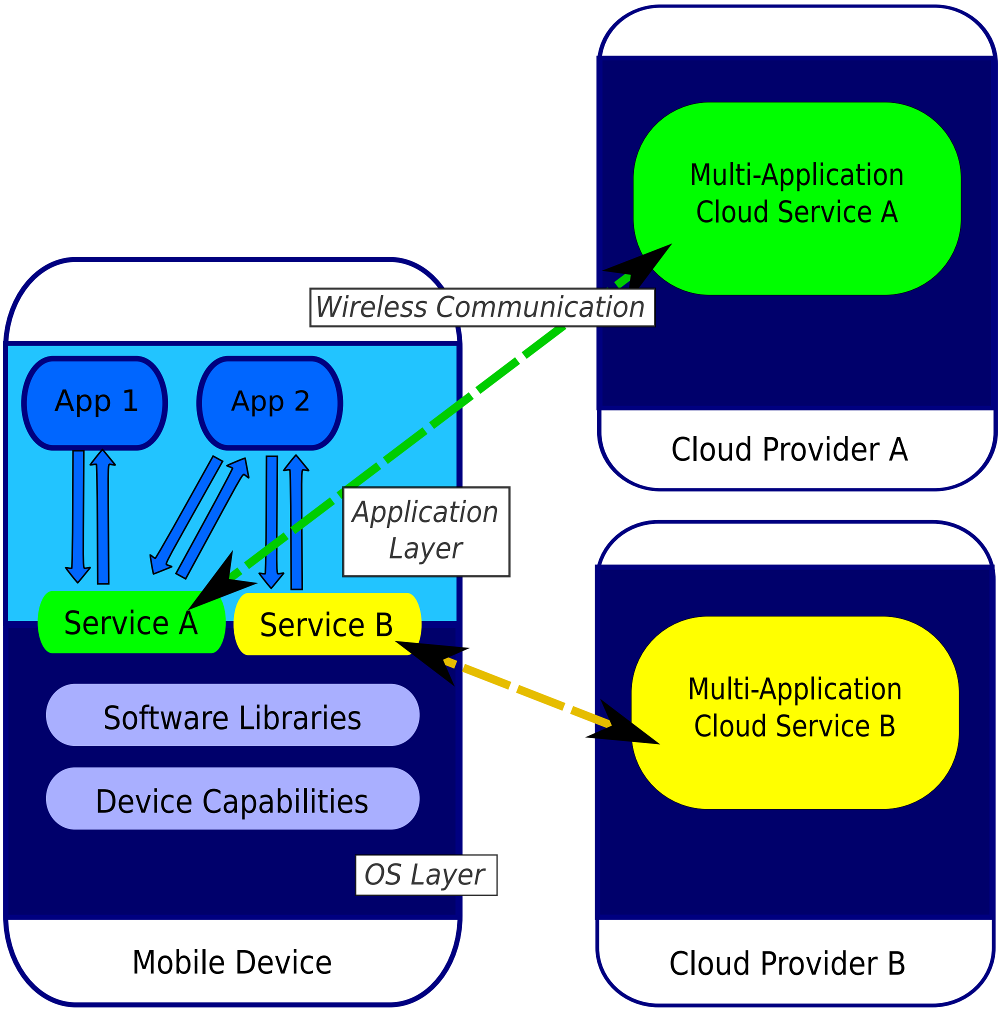

Figure 5.3: A mobile device using two separate multi-application cloud services. Each service provides specific functionality to multiple applications. The software component of the service on the mobile device has elements in both the application and OS layers and communicates directly with the cloud service. In this way, use of the cloud service by applications can be managed to optimize performance and efficiency. Multiple cloud services of this type may be required to provide complete functionality for the mobile device.

In recent scientific literature, various examples of implementation approaches from this category have been presented. Some of the applications of these systems include cloud-based file systems and remote graphical rendering.

Mao et al. have presented the design and implementation of a cloud-oriented file service for mobile devices called *"Wukong"* [112]. This service uses the cloud to augment the storage capacity of mobile devices [112]. Through the software component on the mobile device, this augmented capacity can be used by all applications on the device [112]. Similarly, Nam et al. have used a cloud service to create a *'virtual USB drive'* for mobile devices [113]. Two commercial examples of this type of service are the *"Dropbox"* [114] and *"Box"* [115] services. Both of these provide cloud-based file storage and synchronization functionality to all applications on the mobile device. Kristler and Satyanarayanan have pointed out that when using a network-based file system in the context of portable (mobile) computing, availability is an important consideration [60]. They have described the concept of *'disconnected operation'* in the *"Coda"* file system which enables a client to continue accessing critical data during temporary network failures thus improving the availability of the system [60].

Lu, Li and Shen have proposed a high-level architecture for *'screen virtualization'* or remote screen rendering using a cloud-based service [62]. This service would be used to provide high quality graphically-intense output whilst reducing the computational load on the mobile device [62]. They have explained that this does not imply delegation of all rendering tasks to the cloud but rather a distribution of tasks between the device and the cloud depending on current conditions [62]. Although not explicitly stated, this refers to the use of dynamic partitioning within a multi-application cloud service. Simoens et al. have reviewed and analysed various remote display solutions which use cloud services to provide high quality graphical multimedia content to resource-constrained mobile devices [11].

Since applications do not have direct control over the use of these cloud services, information security is an important consideration in this category. For example, cloud-based file storage services may be used to store sensitive information. Whilst the protection of this information is still primarily the responsibility of the application in which it is used, multi-application cloud services must also provide adequate security measures to protect any information leaving the mobile device. The type of functionality provided by these services determines the type of information involved and therefore defines the required security measures.

Compared to the previous category, failure of a multi-application cloud service poses a higher risk of system unavailability. Such a failure would affect specific functionality across multiple applications on the mobile device. The consequences of service unavailability are dependent on the nature of the service and so availability guarantees should be provided where appropriate. Some services could mitigate this risk through the use of fall-back mechanisms such as the disconnected operation mode in the Coda file system [60].

As shown in **Figure 5.3**, a single mobile device may use multiple cloud services from this category, with each providing different functionality. All interaction with a particular cloud service is managed by the respective software component on the mobile device. This ensures that the functionality provided by the cloud service is allocated to different mobile applications in the optimal manner, thus making the system more efficient from a resource utilization perspective. The streamlining of communication in this architecture improves communication efficiency compared to the previous approach. This also reduces the impact of the system on the mobile network and decreases energy consumption on the mobile device. Another advantage of this type of implementation approach is that it allows for specialization of cloud services such that each cloud service can be optimized to provide a specific type of functionality.

However, there may arise scenarios in which the combination of different cloud services is sub-optimal, especially if they are hosted by different cloud providers. A hypothetical example of this is a system using both a cloud-based file storage service and an enhanced image search service hosted by different providers. Since these two services are not integrated with each other, data can only be transferred between the file service cloud and the image search cloud by passing through the mobile device. This is inefficient because the data is transferred over the wireless network twice. This inefficiency can be overcome through the use of multifunctional cloud resources as explained in the next section.

The multi-application cloud services category represents mid-range implementation approaches for mobile cloud computing. The close integration of these services with the mobile device OS increases their complexity but allows them to provide functionality to multiple applications. In some cases, the use of dynamic partitioning makes these services more adaptable to device heterogeneity and varying network conditions. Information security is an important consideration because these services may involve sensitive information and must therefore ensure that the appropriate security measures are used. The centralized nature of these services increases the risk of system unavailability due to failure of a cloud service. Overall, although they are more complex than the application-specific category, multi-application cloud services provide benefit to more applications and achieve higher levels of efficiency. However, since each service only provides specific functionality, the use of multiple independent services may be required. Although the services will function correctly, this arrangement is less efficient than using a multifunctional cloud resource as explained in the next section.

5.6 Multifunctional Cloud Resources

The multifunctional cloud resources category currently represents the most advanced type of implementation approach for mobile cloud computing. Instead of providing a specific service, the cloud resource provides a dynamic combination of data processing and storage capacity which can be used by any application. As explained in the definition of a cloud resource in **Section 1.3.2**, this processing and storage capacity is significantly greater than that available on the mobile device. The cloud resource also has high-bandwidth fixed communication links to other network end-points which can be used to transfer data on behalf of the mobile device.

In this paradigm, there is a one-to-one relationship between the mobile device and the cloud resource. The full computational capacity of the cloud resource is used to augment the computational capacity of the mobile device. This type of cloud resource is often referred to as a *'surrogate'* [32], [116], [117]. A high-level overview of the architecture of this category of implementation approach is shown in **Figure 5.4**.

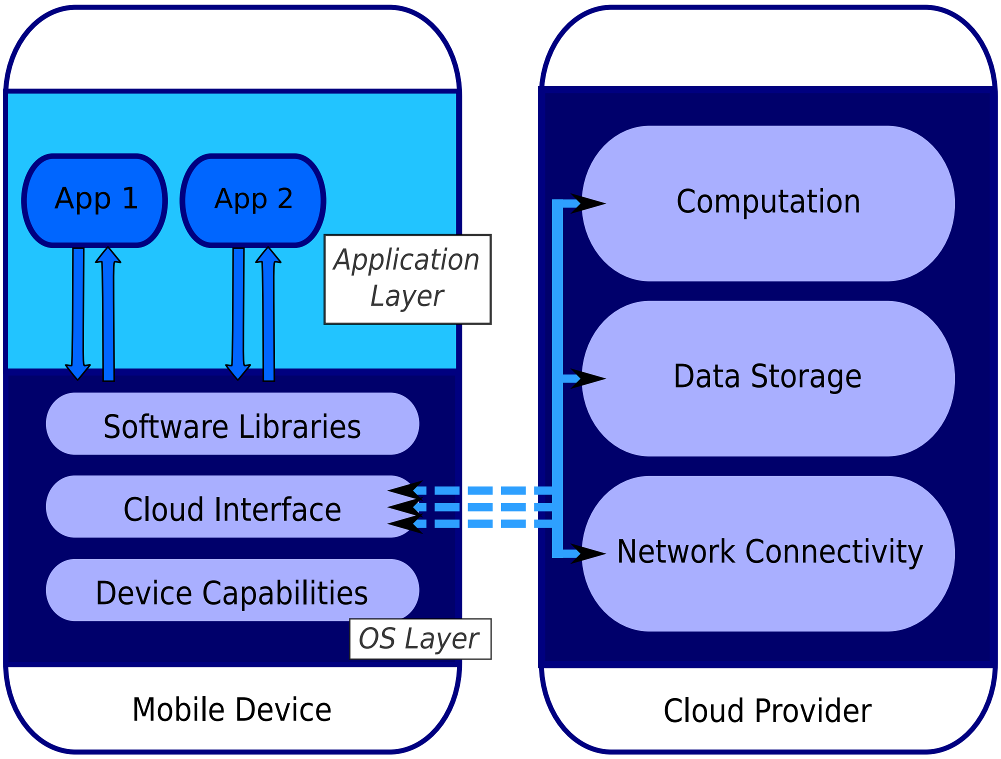

Figure 5.4: A mobile device interacting with a multifunctional cloud resource. This resource provides a dynamic combination of data processing and storage capacity as well as network connectivity. All interaction between the device and the cloud resource is managed and optimized by a software component in the OS layer of the mobile device. The augmented capacity provided by the cloud resource is made available to all applications on the device through this component. In this way, the multifunctional cloud resource is used to enhance the overall computational capacity of the mobile device.

As shown in **Figure 5.4**, the mobile device interacts with the cloud resource through a software component integrated directly into the mobile OS. This low-level integration gives the mobile OS full control over all interactions with the cloud resource. The additional capacity provided by the cloud can be managed by the mobile OS in a similar manner to the device's built-in computational capacity. The mobile OS combines this additional cloud-based computational capacity with the device's existing hardware-based capacity. This augmented capacity is made available to all mobile applications through standard system interfaces so that mobile applications are not aware of the use of the cloud resource. This means that new applications do not need to be specially designed to utilize this augmented capacity and that existing applications can benefit from it without modification. Although they are indicated separately in the figure, the cloud resource provides a dynamic combination of data processing and storage capacity as well as network connectivity.

Implementation approaches in this category have not yet been used in mainstream commercial systems but various examples have been presented in recent scientific literature.

Chun and Maniatis have proposed an architecture for augmenting the computational capacity of smartphones in which mobile applications are partially offloaded to a cloud resource [33]. They have also presented the motivation for the use of dynamic partitioning in such systems and have formalized the dynamic partitioning problem [52]. Most recently, Chun, Maniatis et al. have demonstrated an experimental implementation of this type of approach in the *"CloneCloud"* system [2]. The design of this system is focused on maximizing application performance and minimizing energy consumption [2]. The system does not require any application modification and was evaluated using three types of applications, namely, virus scanning, image search and behaviour profiling applications [2]. Experimental testing has shown that this system can deliver up to 20 times (20x) speed-up and up to 20 times reduction in energy consumption for these applications [2].

Cuervo et al. have demonstrated a similar type of experimental implementation in the *"MAUI"* system [31]. This system makes use of device, network and application profiling in order to dynamically partition applications between the mobile device and the cloud resource [31]. The system has been evaluated using a face recognition application, a graphically intense video game and a chess game. In these applications, the system reduced execution time and energy consumption (by up to an order of magnitude for the face recognition application) [31]. The MAUI system was also used to demonstrate a mobile real-time spoken language translator application which would not otherwise have been possible because its computational requirements exceeded the computational capacity of the mobile device [31].

As demonstrated by the above examples, all approaches in this category use dynamic partitioning. This allows the system to compensate for device heterogeneity and varying network conditions and also allows all applications to utilize the augmented computational capacity provided by the cloud resource. Therefore, this type of approach is highly beneficial to the mobile device as a whole.

From a resource utilization perspective, this is the most efficient approach to mobile cloud computing because all interaction with the cloud is fully managed by the mobile OS. This means that the existing scheduling and resource allocation algorithms which are used for the management of local resources can also be used to manage the cloud resource. This centralized management of the cloud interactions as well as the dynamic partitioning algorithm both require some computational capacity as overhead for the profiling, partitioning and scheduling activities. However, the benefits of these activities exceed their overhead costs as demonstrated by the above example systems.

The use of this centralized management component makes this approach highly efficient from a communication perspective which in turn decreases the impact of the system on the mobile network. The low-level integration with the mobile OS makes it possible to use a comprehensive energy cost model by factoring in device status information such as battery energy and network characteristics such as signal strength.

Information security considerations are similar to the previous approach in that the applications themselves are primarily responsible for securing the information they use. However, the cloud resource must also provide an adequate level of information security. It may be possible to use security considerations as another input to the dynamic partitioning algorithm so that the energy required for additional security services is properly reflected.

One disadvantage of this centralized approach is that it creates a single point of failure. Since unavailability of this system would affect the whole mobile device, it is important to ensure that the system provides acceptable availability guarantees. Since the two examples listed above are experimental systems, this consideration was not addressed in the same level of detail as would be required for a commercial system.

Overall, the multifunctional cloud resource category represents the most advanced type of implementation approach for mobile cloud computing. By providing a dynamic combination of data processing and storage capacity, the cloud resource can augment the overall computational capacity of the mobile device. The use of a centralized management component on the mobile device allows the mobile OS to manage all interaction with the cloud resource. The augmented capacity is made accessible to all mobile applications using standard system interfaces so that no application modification is required. Although this approach has higher computational overhead compared to previous approaches, it is the most efficient in terms of resource utilization. The low-level integration of the management component with the mobile OS allows for the use of detailed energy cost models which improve the energy efficiency of the system. Due to the existence of a single point of failure, availability considerations must be addressed in future implementations. Experimental results showing significant improvements in application performance and energy efficiency have demonstrated that this type of approach represents the greatest potential for future implementations of mobile cloud computing systems.

5.7 Conclusion

This chapter presents an analysis of the various types of implementation approaches for mobile cloud computing and so fulfils the fourth and final requirement of this research as defined in **Section 1.4**. Preceding the core analysis of these implementation approaches, the concepts of mobile grid computing and application partitioning are discussed in order to provide the necessary background information for the analyses in this chapter.

Mobile grid computing involves a group of mobile devices pooling their computational resources in order to perform some computationally intensive operation. These devices all have similar computational capacities and usually interact with each other using short-range wireless links. From an architectural perspective, this is fundamentally different from the mobile cloud computing sub-paradigm in which a resource-constrained mobile device establishes a distributed computing relationship with a remote high-capacity cloud resource. Although the topic of mobile grid computing is beyond the scope of this work, scientific literature from that field has been used to corroborate certain considerations in the analysis framework and to provide examples in some application domains.

The second preliminary section presents important background information about the differences between static and dynamic partitioning of mobile applications. All implementations based on the mobile cloud computing concept use some form of partitioning. In static partitioning, the predefined split between device and cloud makes the system highly deterministic. Static partitioning is relatively simple to implement and can be highly optimized for a particular set of circumstances. However, statically partitioned systems cannot compensate for device heterogeneity or varying network conditions. Dynamic partitioning is more complex to implement and generally requires sophisticated algorithms to partition the software automatically. However, it provides a very high degree of flexibility allowing the system to achieve optimal performance in a variety of different conditions. There have been various efforts to combine the benefits of static and dynamic partitioning into a single approach. There are advantages and disadvantages of both approaches and therefore the choice of partitioning approach ultimately depends on the design and the requirements of the application. This is still an area of ongoing research in which new developments are likely to influence future implementations of mobile cloud computing.

The implementation approaches analysed in this chapter have been selected from recent scientific literature as well as commercial endeavours. A high-level analysis is presented which draws upon aspects of the theoretical analysis framework where appropriate. Based on their architectural characteristics, the implementation approaches discussed in this chapter are grouped into three broad categories, namely, application-specific cloud services, multi-application cloud services and multifunctional cloud resources. By analysing these categories rather than individual applications, it is possible to obtain insight and conclusions which are applicable to a number of different systems.

Application-specific cloud services use the cloud to enhance a single application on the mobile device. These are often proprietary systems in which the cloud service is used for a specific purpose. In this category, static partitioning is used and the mobile application interacts directly with the cloud service. Therefore this type of approach is relatively simple to implement since it does not require modification of the mobile OS. An application-specific cloud service can provide significant benefit to a particular application. However, in a multitasking environment the lack of interaction between different applications using this type of approach results in a silo mentality. Although this silo mentality has advantages in terms of information security and system availability, it is sub-optimal in multitasking environments. Since cloud interaction is not centrally managed by the mobile OS, all active applications on the mobile device could attempt to communicate with their own cloud services simultaneously. This would lead to competition for communication capacity and so decrease the efficiency of the system, especially in the case of multiple connections being established in parallel to the same cloud provider. Overall, this type of approach can be considered to be entry-level mobile cloud computing since it is the simplest to implement and it enhances a specific mobile application. However, since each service only benefits a single application, application-specific cloud services are less beneficial to the mobile device as a whole compared to multi-application cloud services and multifunctional cloud resources.

Multi-application cloud services provide specific functionality to multiple applications on the mobile device. Applications access this functionality through a software component situated on the border between the OS layer and the application layer on the mobile device. The role of this component is to manage and optimize the use of the cloud service. The close integration of this management component with the mobile OS makes this category more complex than the single-application category. Multi-application cloud services use either static or dynamic partitioning depending on the nature of the functionality they provide. Although dynamic partitioning adds complexity to the system, multiple applications benefit from the flexibility it provides. Information security is an important consideration because applications do not have direct control over the use of the cloud service. This type of approach increases the risk of system unavailability since the failure of a single service could affect multiple applications. The use of a centralized management component for each service means that interaction between the various applications and the cloud can be managed and streamlined. This improves communication and energy efficiency on the mobile device and minimizes the impact of the system on the mobile network. However, since the different cloud services are not designed to be interoperable, the only way in which data can move between them is via the mobile device. Overall, this category represents mid-range mobile cloud computing in which each service provides specific functionality to multiple applications on the mobile device. However, in certain situations, this type of approach is less efficient than the use of a multifunctional cloud resource.

Multifunctional cloud resources currently represent the most advanced type of implementation approach for mobile cloud computing. In this category, the cloud resource (surrogate) provides a dynamic combination of data processing and storage capacity as well as network connectivity. The computational capacity of this surrogate is significantly greater than that of the mobile device. The mobile device interacts with the cloud through a management component in the mobile OS layer. This allows the mobile OS to manage cloud capacity as if it were a local resource. The augmented capacity can be utilized by all applications without modification. Although there are no mainstream commercial implementations of this type of approach, experimental systems have been used to demonstrate significant improvements in computational performance and energy efficiency as well as enabling new types of applications. Dynamic partitioning is used to compensate for device and application heterogeneity as well as for fluctuating network conditions. The use of the centralized management component adds some computational overhead but provides significant improvements in computational efficiency. Communication between the mobile device and the cloud can be streamlined to improve communication efficiency and reduce network impact. This low-level integration of the management component with the mobile OS makes it possible to use comprehensive energy cost models. Security requirements are similar to the multi-application category and could be used as an input to the dynamic partitioning algorithm. Availability is an important consideration in this category and must be addressed in future implementations. Overall, this type of approach is the most versatile as it provides augmented multi-functional computational capacity to all applications on the mobile device. Multifunctional cloud resources currently represent the greatest potential for future implementations of mobile cloud computing systems.

All three categories of implementation approaches discussed in this chapter have various advantages and disadvantages. In the short-term, it is likely that all types of approaches will continue to co-exist and will be successfully used in different application domains. However, as the use of mobile cloud computing increases, it is likely that there will ultimately be a shift towards the use of multifunctional cloud resources as explained in the next chapter.

Chapter 6

Conclusions and Recommendations

6.1 Conclusion and Summary

Mobile Computing & Communication Devices (MCCDs) are becoming increasingly prevalent and there has been a significant rise in the use of mobile software applications for these devices (mobile apps). This indicates that there is currently a trend towards the use of mobile computing. However, due to the fundamental constraints of mobility, mobile computing devices will always be resource-constrained compared to non-mobile systems. It has been proposed that the cloud computing paradigm could be used to augment the computational capacity of mobile devices. Using networks such as the internet, it is possible to establish a distributed computing relationship between a mobile device and a remote high-capacity cloud resource. This type of enhanced mobile computing using cloud resources is known as mobile cloud computing.

The overall aim of this minor dissertation is to provide a consolidated review and structured critical analysis of current research and developments within the field of mobile cloud computing. It accomplishes this by satisfying the four major objectives defined in **Section 1.4** which are:

- To investigate the current situation and prevailing trends in mobile computing and present the motivation for enhanced mobile computing.
- To identify and consolidate various considerations from recent literature into a structured theoretical analysis framework for mobile cloud computing.
- To use this framework in the analysis and comparison of selected mobile application domains which could benefit from mobile cloud computing.
- To categorize, analyse and compare various implementation approaches for systems based on the mobile cloud computing paradigm.

In order to obtain sufficient coverage of the major research in this field, a high-level systems approach is used throughout this work. Further information is available in the scientific literature cited as sources in each section. Due to the complexity of this technology and the focus of this minor dissertation, the physical implementation and testing of prototype mobile cloud computing systems is beyond the scope of this research.

Chapter 2 fulfils the first objective by explaining the motivation for the use of enhanced mobile computing.

Data from sources such as the ITU has shown that the number of mobile devices worldwide has increased rapidly over the past ten years. There has also been a significant increase in the use of mobile apps developed for these devices. Both the number of devices and the number of apps are expected to continue increasing thus indicating that there is currently a trend towards the use of mobile computing.

However, the computational capacity of these mobile devices is inherently limited because of the fundamental constraints of mobility. Specifically, the small physical size of the device and its finite energy storage capacity often make it infeasible to use more advanced computational hardware. Small physical size has also been a limiting factor in terms of the UI capabilities of mobile devices.

In comparison to these resource-constrained mobile devices, the emerging cloud computing paradigm provides high-capacity computational resources which are highly elastic in nature. By leveraging economies of scale, cloud providers can use large-scale data centres to increase the efficiency of these systems. The use of virtualization technology allows the computational capacity of the cloud to be provisioned with a high degree of elasticity.

The mobile cloud computing paradigm uses these high-capacity elastic cloud resources to augment the computational capacity of mobile devices through a distributed computing architecture. Three important conditions for the realization of this concept have recently been met. These are the pervasive nature of mobile devices, the broad availability of cloud computing resources and the widespread coverage of mobile communication networks.

These factors have led to the current situation in which there is a growing demand for computational capacity in mobile devices which can be met using cloud resources through the mobile cloud computing paradigm.

Chapter 3 presents a theoretical analysis framework for mobile cloud computing. This framework is a structured consolidation of various considerations identified in recent scientific literature. It therefore fulfils the second objective of this research.

Three important requirements which have guided the design of this framework are that it should be comprehensive in terms of its coverage, flexible in its application and enduring in its relevance.

A high-level qualitative approach is used in the design of this framework. This allows for the inclusion of a broad spectrum of analysis criteria. This qualitative approach also results in more flexible and durable output from the framework since it avoids the use of low-level benchmarks and metrics. Although useful, these metrics and benchmarks are inherently linked to a specific purpose and often become obsolete as technology advances. Since this framework focuses on the fundamental concepts behind mobile cloud computing, it will remain relevant well beyond the current iteration of the technology cycle.

The major considerations in this framework are grouped into the seven logical aspects shown in **Table 6.1**.

Table 6.1: Major aspects of the theoretical analysis framework for mobile cloud computing.

Computational requirements are constituted by the quantity, complexity and type of computational operations performed as well as the volume of data which is processed or stored.
Communication requirements include the quantity of data to be transferred over the network as well as performance metrics such as communication bandwidth and latency.
Mobile network impact analyses the impact of the system on a mobile network and vice-versa. It compares communication requirements to mobile network capabilities.
Energy considerations examine the energy costs and benefits for both the mobile device and the overall system. The system is assessed in terms of energy efficiency.
Information security deals with issues of privacy and security arising from the transfer of data between the mobile device and the cloud over non-private networks.
System availability investigates the risk of system failure and how the consequences of unavailability impact on the usefulness of the system.
Application usability considers any limitations to applications due to the inherently constrained UI capabilities of mobile devices.

Through the process of grouping the considerations into these aspects, various interdependencies between aspects in the framework have been identified as shown in shown in **Table 3.8**. Each of these interdependencies indicates that the major considerations of the relevant aspects have an effect on one another. This highlights the importance of including all these aspects in the analysis of mobile cloud computing systems and thus validates the requirement for the use of a comprehensive analysis framework.

This framework is designed for use in the analysis of mobile cloud computing systems. It also facilitates comparisons between different systems based on particular aspects. Various major considerations identified in this framework can be used in the analysis and comparison of different implementation approaches for mobile cloud computing systems.

Chapter 4 demonstrates the use of the theoretical analysis framework in the analysis of various mobile application domains. In this analysis, mobile apps are grouped into application domains based on their functional characteristics. This allows the domain-specific conclusions and recommendations to be generalized to multiple apps. The application domains discussed in this chapter have been sourced from recent scientific literature as well as existing commercial endeavours and have been selected based on the degree of benefit they would derive from the mobile cloud computing paradigm. However, this cannot be considered an exhaustive list due to the rapid development of new mobile applications.

Some of these application domains have been discussed in recent scientific literature but have been analysed using only a limited subset of criteria. The use of the analysis framework provides a comprehensive qualitative analysis, supported by literature and examples, showing the impact of mobile cloud computing on each aspect of each domain. This has led to the identification of aspects in which the domain would benefit from mobile cloud computing as well as potential issues which must be addressed. It also allows for comparisons between different application domains.

Table 6.2 shows the application domains that have been analysed and highlights the most important conclusions in each domain.

All the application domains featured in this analysis can benefit from the use of mobile cloud computing. Various overall trends have been identified through this analysis. Firstly, domains with high computational require-

Table 6.2: Application domains analysed using the analysis framework.

Mobile scientific computing: Applications in this domain are very computationally-intensive and would not be possible on mobile devices without the use of mobile cloud computing. Their relatively low communication requirements make these systems feasible using mobile networks but usability considerations limit the quantity of data which can be output.
Mobile healthcare services (mHealth): The complex data processing in this domain can be performed in the cloud to increase performance and reduce energy consumption. Although communication requirements are relatively low, information security and system availability requirements are critical in this domain.
Mobile tools for education and training (mLearning): This domain is focused on content and collaboration, both of which can be enhanced through the use of the cloud as a communication proxy. This architecture also increases communication and energy efficiency. Usability is an important consideration.
Advanced mobile Human-Computer Interaction: This domain has very strict communication requirements in terms of latency. However, the use of the cloud can lead to significant improvements in energy efficiency and can enable new types of user-device interaction. Security and availability considerations are very important in this domain.
Entertainment and multimedia services: Although it has a low degree of interactivity, multimedia content involves a very high quantity of data which places a significant load on the mobile network. A cloud-based communication proxy can improve the energy and communication efficiency of applications in this domain.
Mobile gaming applications: Mobile gaming has very high computational and communication requirements. The use of mobile cloud computing can improve energy efficiency and increase output quality of applications in this domain up to the maximum limit imposed by the mobile device UI.

ments derive the most direct benefit from the use of a cloud resource. Communication requirements are often influenced by bandwidth and latency considerations and in turn affect the mobile network impact aspect. Both computational and communication requirements have a direct effect on energy consumption as shown by the energy cost model. Security and availability considerations vary widely but still need to be taken into account. In most cases, application usability acts as a limiting factor and is also therefore an important consideration in the analysis.

The analysis of these application domains demonstrates the benefits of using the theoretical analysis framework. This has resulted in various conclusions and recommendations specific to each domain. It has also allowed for qualitative comparisons between different domains based on specific aspects. The information obtained from this analysis can be fed back into the design process in order to enhance future applications using the mobile cloud computing paradigm.

Based on the analyses in this chapter, a generic analysis process has been defined. Although the exact process will depend on the system being analysed, the following overarching stages will be present in all analyses:

- System Identification
- System Verification
- Mobile-Cloud Relationship Analysis
- Analysis of Aspects
- Analysis of Interdependencies
- Conclusion of Analysis

Using this process, the framework defined in **Chapter 3** can be used in the analysis of other systems based on the mobile cloud computing paradigm.

Chapter 5 presents an analysis of the various types of implementation approaches for mobile cloud computing and so fulfils the fourth and final requirement of this research.

Preceding the core analysis in this chapter, the concepts of mobile grid computing and application partitioning are discussed in order to provide the required background information. The mobile grid computing sub-paradigm involves a group of mobile devices pooling their computational resources via short-range wireless networks. From an architectural perspective, this is fundamentally different from the mobile cloud computing paradigm in which a resource-constrained mobile device establishes a distributed computing relationship with a remote high-capacity cloud resource. Although mobile grid computing is beyond the scope of this research, various concepts and ideas in this research have been sourced from mobile grid computing literature. The second preliminary section explains the concepts of static and dynamic partitioning of mobile applications. Static partitioning is simple to implement and highly deterministic. However, it is inflexible and less efficient than dynamic partitioning. Dynamic partitioning is more complex to implement but allows the system to adapt and compensate for device heterogeneity and varying mobile network conditions. Application partitioning is a field of ongoing research in which new developments will have an impact on mobile cloud computing.

The implementation approaches analysed in this chapter have been sourced from recent scientific literature as well as commercial endeavours. A high-level analysis is presented which draws upon aspects of the theoretical analysis framework where appropriate. Based on their architectural characteristics, the implementation approaches are grouped into three broad categories in order to obtain insight and conclusions which are applicable to a number of different systems.

Application-specific Cloud Services use the cloud to enhance a single application on the mobile device. In this category, each mobile app interacts directly with the cloud service. This type of approach is the simplest to implement because it does not require modification of the mobile OS yet it can still provide significant benefit to the specific application. However, in a multitasking environment the lack of interaction between different applications using this type of approach results in a type of silo mentality. Since cloud interaction is not managed by the mobile OS, there could be competition for limited communication resources and multiple connections to the same cloud provider thus reducing the efficiency of the system. Overall, this type of approach can be considered to be entry-level mobile cloud computing.

Multi-application Cloud Services provide specific functionality to multiple applications on the mobile device. In this category, either static or dynamic partitioning schemes are used depending on the type of functionality provided. Applications access this functionality through a centralized management component situated on the border between the OS layer and the application layer on the mobile device. The close integration of this component with the mobile OS makes this category more complex than single-application services. The use of centralized management components for each service results in streamlined communication, reduced network impact and improved energy efficiency. However, a single device may use multiple services from this category which are not integrated with each other thus leading to reduced communication efficiency in some scenarios. Overall, this category represents mid-range mobile cloud computing.

Multifunctional Cloud Resources currently represent the most advanced type of implementation approach for mobile cloud computing. In this category, the cloud resource (surrogate) provides a dynamic combination of data processing and storage capacity as well as external network connectivity. The mobile device interacts with the cloud through a management component in the mobile OS allowing this capacity to be managed by the OS as if it were a local resource. This augmented capacity is made accessible to all applications without modification. Dynamic partitioning is used in order to utilize the flexibility provided by multifunctional cloud resources. Experimental systems have shown that this type of approach can provide significant improvements in computational performance and energy efficiency as well as enabling new types of applications. The use of the centralized management component on the mobile device achieves highly streamlined communication and minimizes energy consumption by enabling comprehensive energy cost models. Overall, multifunctional cloud resources are the most versatile systems and represent the greatest potential for future implementations of mobile cloud computing.

All three categories of implementation approaches have various advantages and disadvantages. In the short-term, it is likely that all three types of approaches will continue to co-exist and will be successfully used in different application domains. As the use of mobile cloud computing increases, it is likely that the use of multifunctional cloud resources will become significantly more prevalent.

Overall, through the development and application of the new analysis framework, this work provides a consolidated review and structured critical analysis of the current research and developments in the field of mobile cloud computing.

6.2 Predicted Future Trends

This section is a discussion of emerging and predicted future trends within the field of mobile cloud computing. Although these predictions are based on recent scientific literature and business intelligence reports, they are still speculative in nature. Although they are not used as the basis for any analysis, these predicted future trends are discussed in order to demonstrate that this area of technology is currently expanding from both research and commercial perspectives. These predictions therefore serve to further highlight the importance of this work.

6.2.1 Mobile Network Operator Involvement

The first predicted future trend is that Mobile Network Operators (MNOs) will become more involved in the implementation of mobile cloud computing systems. In this research, as in most scientific literature, it is assumed that the mobile network provides only a data communication link between the mobile device and the cloud resource. There is a compelling business case for MNOs to diversify so as to avoid being reduced to simple *'bit-pipes'* but that is beyond the scope of this research. From a technological perspective, the involvement of MNOs could benefit the mobile cloud computing paradigm. As explained by Nkosi and Mekuria [73] as well as Klein et al. [42], technologies such as the Internet Protocol Multimedia Subsystem (IMS) could be used to provide additional functionality such as authentication of mobile devices. IMS platforms are typically implemented and managed by MNOs. A further possibility is that MNOs will capitalize on the need for low-latency connections between the mobile device and cloud resource. This would be achieved by closer integration of the cloud provider and the MNO. For example, the *"Amazon Web Services Direct Connect"* service establishes high-performance dedicated network connections between customer networks and Amazon's cloud computing data centres [118]. In some cases, MNOs could become cloud providers leading to cellular mobile cloud computing as discussed in **Section 6.3.4**.

6.2.2 Increase in Mobile Cloud Computing

The second predicted trend is that the use of mobile cloud computing will continue to increase. As explained in **Chapter 2**, the prerequisite conditions have been met and mobile application developers are beginning to take advantage of this technology. This will increase the total volume of communication between mobile devices and the cloud. Fortunately, it is expected that additional electromagnetic spectrum will be allocated for mobile communication purposes as a result of the switch from analogue to digital television broadcast. This is generally referred to as the *'Digital Dividend'*. The report on the Digital Dividend from the GSM Association (GSMA) has listed various examples of new and enhanced mobile broadband services which would benefit from this additional capacity [119]. Many of the services listed in this report will involve mobile cloud computing, including "Medical diagnosis performed remotely", "Mobile gaming" and "Remote data processing" [119]. The impact on the mobile network will still be an important consideration in the design of these systems. It is also likely that each cloud provider will begin to host multiple services. Therefore, the need to maximize communication efficiency will cause a shift away from single-application implementation approaches towards the use of multifunctional cloud resources.

6.3 Recommendations for Future Work

Based on the conclusions of this research and the predicted future trends in **Section 6.2**, this section presents recommendations for future work in the field of mobile cloud computing.

6.3.1 Further Application of the Analysis Framework

Future use of the analysis framework defined in **Chapter 3** will serve to further improve and enhance this framework. Potential opportunities for future improvement include the development of a weighting system for the considerations within each aspect. This would allow for quantitative comparisons between aspects. The development of such a system would be based on a significant number of analyses in which the framework is used. Furthermore, all future use serves as validation of the utility of this framework.

6.3.2 Multifunctional Cloud Resources

As described in **Section 5.6**, multifunctional cloud resources currently represent the most promising type of implementation approach for mobile cloud computing. Some preliminary experimental work has been conducted in the implementation of systems of this type but there is still scope for further investigation since there are a number of research opportunities in this field. These include a comprehensive design and evaluation of the cloud interface component for the mobile OS as well as the standardization of the communication protocol between the mobile device and the cloud resource.

6.3.3 Initial Association Phase

A related research opportunity involves the initial association phase between the mobile device and a cloud resource. In this research, as in most literature, it has been assumed that the mobile device has been successfully associated with a cloud resource. This process involves the discovery of the available cloud resources, selection of the best option and mutual authentication between the mobile device and the cloud resource. This is particularly important in the case of geographically localized cloud resources such as the *'cloudlets'* described by Satyanarayanan et al. [30].

6.3.4 Cellular Mobile Cloud Computing

If geographically localized cloud resources are to be used, the geographical distance from which these resources can be accessed will inherently be limited. Since this system is designed for mobile users, it must allow the user to maintain an ongoing session whilst changing geographical location. Therefore, it will be necessary to perform hand-overs between geographically localized cloud resources. This architecture is very similar to that used in cellular communication networks thus leading to the idea of cellular mobile cloud computing. Given that MNOs are likely to become more involved in mobile cloud computing, this could lead to some form of amalgamation between mobile cellular communication and cellular cloud computing systems which presents many possibilities for future research.

6.3.5 Devices with Limited User Interface

In terms of the theoretical analysis framework defined in this work, further analysis could be carried out to determine how this framework could be adapted or extended for the analysis of smart devices with minimal UI capabilities. These new smart devices are an integral part of the *'internet of things'* concept which is predicted to become a reality in the near future. Ericsson has anticipated that there will be more than 50 billion network-connected devices by the year 2020 [120]. Although they lack direct interaction with the user, these network-connected devices will respond to inputs such as environmental stimuli or remote control. The area of Machine-to-Machine communication (M2M) is already developing rapidly and could possibly benefit from the mobile cloud computing paradigm.

6.3.6 Cloud Perspective of Mobile Cloud Computing

Another future research possibility is to investigate mobile cloud computing from the cloud's perspective. This research has modelled the cloud as a resource which provides a dynamic combination of data processing and storage capacity. Although this model is sufficiently accurate for the purpose of this research, it does not explain how the cloud-based components of these systems would be implemented. There is scope for the design of new software systems for the cloud to facilitate interaction with mobile devices. These systems would be responsible for the allocation of cloud resources between different mobile devices as well as the management of connections between devices and the cloud. It may be possible to optimize cloud systems for use in mobile cloud computing using new software architectures and enhanced virtualization capabilities.

6.3.7 Mutually Beneficial Distributed Computing

The final recommendation for future work in this field relates to the development of a mutually beneficial distributed computing relationship between the mobile device and the cloud. The fundamental design of the current relationship is to use the cloud to enhance the mobile device. However, it may be possible to balance this situation by also using mobile devices to benefit the cloud. The primary advantage of mobile devices is the significant quantity of these devices which are geographically distributed throughout the world. Each device usually features a number of sensors which capture input data from the environment. By amalgamating this data from multiple different sources, the cloud could perform useful computation and analysis using the principle of crowd sourcing. An example of this would be to use mobile devices to gather weather data in specific regions to supplement weather simulations being performed in the cloud. Further examples are provided by two hypothetical future scenarios proposed by Satyanarayanan [121]. This is essentially an interconnection of the vast computational resources of the cloud with the global decentralized network of sensor-enabled mobile devices and as such is an important opportunity for future work.

6.3.8 Summary of Recommendations for Future Work

Based on this research, various opportunities for future work have been identified. Firstly, further use of the analysis framework will result in improvements to the framework and serve as validation of the utility of this work. Secondly, more extensive experimentation and testing can be conducted in the area of multifunctional cloud resources. This includes the design of the cloud interface software component and the standardization of the communication protocol. The third research opportunity involves the initial association phase between the mobile device and the cloud, including service discovery, selection and authentication. The use of geographically localized cloud resources also presents various research opportunities, particularly in dealing with user mobility. Due to the architectural similarities, this could lead to an amalgamation of mobile cellular communication and cellular cloud computing systems. Another possibility for future work is to investigate how the theoretical analysis framework can be applied to mobile devices with minimal UI capabilities. This includes cloud-enhanced M2M communication and forms part of the *'internet of things'* concept. The sixth proposal for future work is to conduct an investigation similar to this from the cloud's perspective. It may be possible to optimize cloud systems for use in mobile cloud computing using new software architectures and enhanced virtualization capabilities. The final recommendation is to investigate the possibility of a mutually-beneficial distributed computing relationship between mobile devices and cloud resources. In such a relationship, the cloud would also benefit from the geographically distributed nature of mobile devices and the abundance of sensors present on each device. This type of system is essentially an interconnection of the vast computational resources of the cloud with the extensive globally decentralized network of cloud-enhanced sensor-enabled mobile devices.

Bibliography

[1] K. Kumar and Y.-H. Lu, "Cloud Computing for Mobile Users: Can Offloading Computation Save Energy?" *Computer*, vol. 43, no. 4, pp. 51–56, Apr. 2010.

[2] B.-G. Chun, S. Ihm, P. Maniatis, M. Naik, and A. Patti, "CloneCloud: Elastic Execution between Mobile Device and Cloud," in *Proceedings of the sixth conference on Computer systems - EuroSys '11*, ser. EuroSys '11. New York, New York, USA: ACM Press, 2011, p. 301.

[3] E. E. Marinelli, "Hyrax: Cloud Computing on Mobile Devices using MapReduce," Masters thesis, Carnegie-Mellon University, Pittsburgh, PA, 2009.

[4] International Data Corporation (IDC), "Worldwide Quarterly Mobile Phone Tracker," 2011, [Online]. Available:
http://www.idc.com/getdoc.jsp?containerId=prUS22679411
Last Accessed: 17 August 2011.

[5] International Data Corporation (IDC), "Worldwide Quarterly PC Tracker," 2011, [Online]. Available:
http://www.idc.com/getdoc.jsp?containerId=prUS22653511
Last Accessed: 17 August 2011.

[6] International Telecommunication Union (ITU), "Global ICT Developments," 2011, [Online]. Available:

http://www.itu.int/ITU-D/ict/statistics/
Last Accessed: 4 December 2011.

[7] International Data Corporation (IDC), "Worldwide and U.S. Mobile Applications, Storefronts, and Developer 2010-2014 Forecast and Year-End 2010 Vendor Shares: The 'Appification' of Everything," 2010, [Online]. Available:
http://www.idc.com/getdoc.jsp?containerId=225668
Last Accessed: 4 December 2011.

[8] Gartner Inc., "Forecast: Mobile Application Stores, Worldwide, 2008-2014," 2011, [Online]. Available:
http://www.gartner.com/resId=1498914
Last Accessed: 4 December 2011.

[9] National Institute of Standards and Technology (NIST), "The NIST Definition of Cloud Computing," Std. 800-145, 2011.

[10] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "Above the Clouds: A Berkeley View of Cloud Computing," EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2009-28, 2009.

[11] P. Simoens, F. De Turck, B. Dhoedt, and P. Demeester, "Remote display solutions for mobile cloud computing," *Computer*, no. 99, pp. 1–1, 2011.

[12] In-Stat, "Theres an App for That: Smartphone Application Market Overcrowded but Downloads Continue to Rise," 2011, [Online]. Available:
http://www.instat.com/abstract.asp?id=231&SKU=IN1104930MCM
Last Accessed: 4 December 2011.

[13] Juniper Research, "Mobile App Stores Business Models, Strategies & Market Segmentation 2010-2015," 2010, [Online]. Available:

http://juniperresearch.com/reports/mobile_app_stores
Last Accessed: 4 December 2011.

[14] M. Satyanarayanan, "Fundamental challenges in mobile computing," in *Proceedings of the fifteenth annual ACM symposium on Principles of distributed computing - PODC '96*. New York, USA: ACM Press, May 1996, pp. 1–7.

[15] R. Rosenthal and M. R. DiMatteo, "META -ANALYSIS : Recent Developments in Quantitative Methods for Literature Reviews," *Annual review of psychology*, vol. 59, pp. 59–82, 2001.

[16] M. Weiser, "Some computer science issues in ubiquitous computing," *Communications of the ACM*, vol. 36, no. 7, pp. 75–84, Jul. 1993.

[17] G. Forman and J. Zahorjan, "The challenges of mobile computing," *Computer*, vol. 27, no. 4, pp. 38–47, Apr. 1994.

[18] International Telecommunication Union (ITU), "World Telecommunication/ICT Indicators Database," 2011, [Online]. Available:
http://www.itu.int/ITU-D/ict/publications/world/world.html
Last Accessed: 4 December 2011.

[19] TouchType Ltd., "What do smartphone users really want?" 2011, [Online]. Available:
http://www.swiftkey.net/what-do-smartphone-users-really-want
Last Accessed: 15 November 2011.

[20] National Research Council Committee on Electric Power for the Dismounted Soldier, *Energy-Efficient Technologies for the Dismounted Soldier*. The National Acadamies Press, 1997.

[21] IEEE Standards Association, "IEEE Launches Pioneering Cloud Computing Initiative - Press Release," 2011, [Online]. Available:
http://standards.ieee.org/news/2011/cloud.html
Last Accessed: 15 August 2011.

[22] J. Hamilton, "Cloud-Computing Economies of Scale (AWS Workshop on Genomics and Cloud Computing)," 2010, [Online]. Available: http://www.mvdirona.com/jrh/TalksAndPapers/JamesHamilton_GenomicsCloud20100608.pdf
Last Accessed: 20 December 2011.

[23] Google Inc., "Google App Engine," 2011, [Online]. Available: http://code.google.com/appengine/
Last Accessed: 26 November 2011.

[24] Amazon Web Services, "Amazon Elastic Compute Cloud (Amazon EC2)," 2011, [Online]. Available: http://aws.amazon.com/ec2/
Last Accessed: 26 November 2011.

[25] Microsoft Corp., "Cloud Computing & Storage from Microsoft," 2011, [Online]. Available: http://www.microsoft.com/cloud
Last Accessed: 26 November 2011.

[26] S. Srikantaiah, A. Kansal, and F. Zhao, "Energy aware consolidation for cloud computing," in *Proceedings of the 2008 conference on Power aware computing and systems*, Dec. 2008, p. 10.

[27] Y. Tian, B. Song, and E.-N. Huh, "Towards the Development of Personal Cloud Computing for Mobile Thin-Clients," in *2011 International Conference on Information Science and Applications*. IEEE, Apr. 2011, pp. 1–5.

[28] A. Messer, I. Greenberg, P. Bernadat, D. Milojicic, D. Chen, T. Giuli, and X. Gu, "Towards a distributed platform for resource-constrained devices," in *Proceedings 22nd International Conference on Distributed Computing Systems*. IEEE Comput. Soc, 2002, pp. 43–51.

[29] R. Balan, J. Flinn, M. Satyanarayanan, S. Sinnamohideen, and H.-I. Yang, "The case for cyber foraging," in *Proceedings of the 10th workshop on ACM SIGOPS European workshop: beyond the PC - EW10*. New York, New York, USA: ACM Press, Jul. 2002, p. 87.

[30] M. Satyanarayanan, P. Bahl, R. Cáceres, and N. Davies, "The Case for VM-Based Cloudlets in Mobile Computing," *IEEE Pervasive Computing*, vol. 8, no. 4, pp. 14–23, Oct. 2009.

[31] E. Cuervo, A. Balasubramanian, D.-k. Cho, A. Wolman, S. Saroiu, R. Chandra, and P. Bahl, "MAUI: Making Smartphones Last Longer with Code Offload," in *Proceedings of the 8th international conference on Mobile systems, applications, and services - MobiSys '10*. New York, New York, USA: ACM Press, Jun. 2010, p. 49.

[32] Y. Kun, O. Shumao, and C. Hsiao-Hwa, "On effective offloading services for resource-constrained mobile devices running heavier mobile Internet applications," *IEEE Communications Magazine*, vol. 46, no. 1, pp. 56–63, Jan. 2008.

[33] B.-G. Chun and P. Maniatis, "Augmented smartphone applications through clone cloud execution," in *Proceedings of the 12th conference on Hot topics in operating systems*, ser. HotOS'09. Berkeley: USENIX Association, 2009, pp. 8–8.

[34] D. Narayanan, J. Flinn, and M. Satyanarayanan, "Using history to improve mobile application adaptation," in *Proceedings Third IEEE Workshop on Mobile Computing Systems and Applications*. IEEE Comput. Soc, 2000, pp. 31–40.

[35] J. Flinn, D. Narayanan, and M. Satyanarayanan, "Self-tuned remote execution for pervasive computing," in *Proceedings Eighth Workshop on Hot Topics in Operating Systems*. IEEE Comput. Soc, 2001, pp. 61–66.

[36] G. M. Amdahl, "Validity of the single processor approach to achieving large scale computing capabilities," in *Proceedings of the April 18-20, 1967, spring joint computer conference on - AFIPS '67 (Spring)*. New York, New York, USA: ACM Press, Apr. 1967, p. 483.

[37] Amazon Web Services, "Amazon Simple Storage Service (Amazon S3)," 2011, [Online]. Available: http://aws.amazon.com/s3/ Last Accessed: 30 November 2011.

[38] C. Wang and Z. Li, "Parametric analysis for adaptive computation offloading," in *Proceedings of the ACM SIGPLAN 2004 conference on Programming language design and implementation - PLDI '04*, vol. 39, no. 6. New York, New York, USA: ACM Press, Jun. 2004, p. 119.

[39] R. Wolski, S. Gurun, C. Krintz, and D. Nurmi, "Using bandwidth data to make computation offloading decisions," in *2008 IEEE International Symposium on Parallel and Distributed Processing*. IEEE, Apr. 2008, pp. 1–8.

[40] N. Tolia, D. Andersen, and M. Satyanarayanan, "Quantifying Interactive User Experience on Thin Clients," *Computer*, vol. 39, no. 3, pp. 46–52, Mar. 2006.

[41] I. Kelenyi and J. K. Nurminen, "CloudTorrent - Energy-Efficient BitTorrent Content Sharing for Mobile Devices via Cloud Services," in *2010 7th IEEE Consumer Communications and Networking Conference*. IEEE, Jan. 2010, pp. 1–2.

[42] A. Klein, C. Mannweiler, J. Schneider, and H. D. Schotten, "Access Schemes for Mobile Cloud Computing," in *2010 Eleventh International Conference on Mobile Data Management*. IEEE, May 2010, pp. 387–392.

[43] M. Satyanarayanan, "Avoiding Dead Batteries," *IEEE Pervasive Computing*, vol. 4, no. 1, pp. 2–3, Jan. 2005.

[44] J. Flinn and M. Satyanarayanan, "Energy-aware adaptation for mobile applications," *ACM SIGOPS Operating Systems Review*, vol. 33, no. 5, pp. 48–63, Dec. 1999.

[45] E. Lagerspetz and S. Tarkoma, "Mobile search and the cloud: The benefits of offloading," in *2011 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*. IEEE, Mar. 2011, pp. 117–122.

[46] A. P. Miettinen and J. K. Nurminen, "Energy efficiency of mobile clients in cloud computing," *Energy*, p. 4, 2010.

[47] S. Cho and R. Melhem, "Corollaries to Amdahl's Law for Energy," *IEEE Computer Architecture Letters*, vol. 7, no. 1, pp. 25–28, Jan. 2008.

[48] F.-S. Chu, K.-C. Chen, and C.-M. Cheng, "Toward green cloud computing," in *Proceedings of the 5th International Confernece on Ubiquitous Information Management and Communication - ICUIMC '11*, ser. ICUIMC '11. New York, New York, USA: ACM Press, 2011, p. 1.

[49] International Organization for Standardization (ISO), "Information Technology - Security Techniques - Code of Practice for Information Security Management," Tech. Rep., 2005, [Online]. Available: http://www.iso.org/iso/catalogue_detail?csnumber=50297 Last Accessed: 4 December 2011.

[50] U.S. Department of Health & Human Services (HHS), "The Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy and Security Rules," 1996, [Online]. Available: http://www.hhs.gov/ocr/privacy/hipaa/understanding/index.html Last Accessed: 4 December 2011.

[51] Amazon Web Services, "Creating HIPAA-Compliant Medical Data Applications With AWS," Tech. Rep., 2009, [Online]. Available: http://awsmedia.s3.amazonaws.com/AWS_HIPAA_Whitepaper_Final.pdf Last Accessed: 4 December 2011.

[52] B.-G. Chun and P. Maniatis, "Dynamically partitioning applications between weak devices and clouds," in *Proceedings of the 1st ACM*

Workshop on Mobile Cloud Computing & Services Social Networks and Beyond - MCS '10, ser. MCS '10. New York, New York, USA: ACM Press, 2010, pp. 1–5.

[53] National Institute of Standards and Technology (NIST), "Guide to Storage Encryption Technologies for End User Devices," Tech. Rep., 2007.

[54] M. Brenner, J. Wiebelitz, G. von Voigt, and M. Smith, "Secret program execution in the cloud applying homomorphic encryption," in *5th IEEE International Conference on Digital Ecosystems and Technologies (IEEE DEST 2011)*. IEEE, May 2011, pp. 114–119.

[55] W. Itani, A. Kayssi, and A. Chehab, "Energy-efficient incremental integrity for securing storage in mobile cloud computing," in *2010 International Conference on Energy Aware Computing*. IEEE, Dec. 2010, pp. 1–2.

[56] A. Kejariwal, S. Gupta, A. Nicolau, N. Dutt, and R. Gupta, "Energy efficient watermarking on mobile devices using proxy-based partitioning," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 14, no. 6, pp. 625–636, Jun. 2006.

[57] S. Wang and X. S. Wang, "In-Device Spatial Cloaking for Mobile User Privacy Assisted by the Cloud," in *2010 Eleventh International Conference on Mobile Data Management*. IEEE, 2010, pp. 381–386.

[58] Amazon Web Services, "Amazon S3 Service Level Agreement," 2007, [Online]. Available: http://aws.amazon.com/s3-sla/
Last Accessed: 4 December 2011.

[59] M. Satyanarayanan, J. Kistler, P. Kumar, M. Okasaki, E. Siegel, and D. Steere, "Coda: a highly available file system for a distributed workstation environment," *IEEE Transactions on Computers*, vol. 39, no. 4, pp. 447–459, Apr. 1990.

[60] J. J. Kistler and M. Satyanarayanan, "Disconnected operation in the Coda File System," *ACM Transactions on Computer Systems*, vol. 10, no. 1, pp. 3–25, Feb. 1992.

[61] L. Chittaro, "Visualizing Information on Mobile Devices," *Computer*, vol. 39, no. 3, pp. 40–45, Mar. 2006.

[62] Y. Lu, S. Li, and H. Shen, "Virtualized Screen: A Third Element for Cloud-Mobile Convergence," *IEEE Multimedia*, vol. 18, no. 2, pp. 4–11, Feb. 2011.

[63] E. Rukzio, P. Holleis, and H. Gellersen, "Personal Projectors for Pervasive Computing," *IEEE Pervasive Computing*, no. 99, pp. 1–1, 2011.

[64] L. Terrenghi, T. Lang, and B. Lehner, "Elastic Mobility : Stretching Interaction," *Gesture*, p. 1, 2009.

[65] M. Masinde and A. Bagula, "A framework for predicting droughts in developing countries using sensor networks and mobile phones," in *Proceedings of the 2010 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on - SAICSIT '10.* New York, New York, USA: ACM Press, Oct. 2010, pp. 390–393.

[66] Wolfram Alpha LLC, "Wolfram Alpha: Mobile & Tablet Apps," 2011, [Online]. Available: http://products.wolframalpha.com/mobile/ Last Accessed: 30 October 2011.

[67] Amazon Web Services, "High Performance Computing (HPC) on AWS," 2011, [Online]. Available: http://aws.amazon.com/hpc-applications/ Last Accessed: 29 November 2011.

[68] J. Dean and S. Ghemawat, "MapReduce: simplified data processing on large clusters," in *Proceedings of the 6th conference on Symposium*

on *Opearting Systems Design & Implementation - Volume 6*, Dec. 2004, p. 10.

[69] M. Masinde, A. Bagula, and V. Ndegwa, "MobiGrid: A middleware for integrating mobile phone and grid computing," in *2010 International Conference on Network and Service Management*. IEEE, Oct. 2010, pp. 523–526.

[70] K. Hameed, "The application of mobile computing and technology to health care services," *Telematics and Informatics*, vol. 20, no. 2, pp. 99–106, May 2003.

[71] U. Varshney, "Pervasive healthcare," *Computer*, vol. 36, no. 12, pp. 138–140, Dec. 2003.

[72] D. N. Breslauer, R. N. Maamari, N. a. Switz, W. a. Lam, and D. a. Fletcher, "Mobile phone based clinical microscopy for global health applications." *PloS one*, vol. 4, no. 7, p. e6320, Jan. 2009.

[73] M. Nkosi and F. Mekuria, "Cloud Computing for Enhanced Mobile Health Applications," in *2010 IEEE Second International Conference on Cloud Computing Technology and Science*. IEEE, Nov. 2010, pp. 629–633.

[74] D. B. Hoang and L. Chen, "Mobile Cloud for Assistive Healthcare (MoCAsH)," in *2010 IEEE Asia-Pacific Services Computing Conference*. IEEE, Dec. 2010, pp. 325–332.

[75] C. Doukas, T. Pliakas, and I. Maglogiannis, "Mobile healthcare information management utilizing Cloud Computing and Android OS." *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, vol. 2010, pp. 1037–40, Jan. 2010.

[76] U. Farooq, W. Schafer, M. Rosson, and J. Carroll, "M-Education: bridging the gap of mobile and desktop computing," in *Proceedings.*

IEEE International Workshop on Wireless and Mobile Technologies in Education. IEEE Comput. Soc, 2002, pp. 91–94.

[77] YouTube LLC, "YouTube for Education," 2011, [Online]. Available: http://www.youtube.com/education
Last Accessed: 14 December 2011.

[78] J. Joubert, "Personal communication," 2011.

[79] M. A. H. Masud, J. Yong, and X. Huang, "Enhanced M-Learning with cloud computing: The Bangladesh case," in *Proceedings of the 2011 15th International Conference on Computer Supported Cooperative Work in Design (CSCWD).* IEEE, Jun. 2011, pp. 735–741.

[80] A. Tanwer, A. Tayal, M. Hussain, and P. Reel, "Thin apps store for smart phones based on private cloud infrastructure," in *Kaleidoscope: Beyond the Internet? - Innovations for Future Networks and Services, 2010 ITU-T*, 2010, pp. 1–6.

[81] V. Glavinic, S. Ljubic, and M. Kukec, "A Holistic Approach to Enhance Universal Usability in m-Learning," in *2008 The Second International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies.* IEEE, Sep. 2008, pp. 305–310.

[82] Y. Niu, X. Li, X. Meng, J. Sun, and H. Dong, "A Constraint-based User Interface Design Method for Mobile Computing Devices," in *2006 First International Symposium on Pervasive Computing and Applications.* IEEE, Aug. 2006, pp. 343–347.

[83] I. Kondratova and I. Goldfarb, "M-learning: Overcoming the Usability Challenges of Mobile Devices," in *International Conference on Networking, International Conference on Systems and International Conference on Mobile Communications and Learning Technologies (ICNICONSMCL'06).* IEEE, 2006, pp. 223–223.

[84] Apple Inc., "Siri - Frequently Asked Questions," 2011, [Online]. Available: http://www.apple.com/iphone/features/siri-faq.html Last Accessed: 29 November 2011.

[85] M. Ghaziasgar and J. Connan, "Re-Designing the iSign Front-End Interface," in *Southern Africa Telecommunications Networks and Applications Conference (SATNAC)*, 2011.

[86] P. Angin, B. Bhargava, and S. Helal, "A Mobile-Cloud Collaborative Traffic Lights Detector for Blind Navigation," in *2010 Eleventh International Conference on Mobile Data Management*. IEEE, 2010, pp. 396–401.

[87] Y.-J. Hong, K. Kumar, and Y.-H. Lu, "Energy efficient content-based image retrieval for mobile systems," in *2009 IEEE International Symposium on Circuits and Systems*. IEEE, May 2009, pp. 1673–1676.

[88] H. A. Lagar-Cavilla, N. Tolia, E. de Lara, M. Satyanarayanan, and D. O'Hallaron, "Interactive resource-intensive applications made easy," in *Proceedings of the ACM/IFIP/USENIX 2007 International Conference on Middleware*, Nov. 2007, pp. 143–163.

[89] Cisco Systems Inc., "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2010 - 2015," Tech. Rep., 2011, [Online]. Available: http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.pdf Last Accessed: 4 December 2011.

[90] H. Shen, Z. Pan, H. Sun, Y. Lu, and S. Li, "A proxy-based mobile web browser," in *Proceedings of the international conference on Multimedia - MM '10*, ser. MM '10. New York, New York, USA: ACM Press, 2010, p. 763.

[91] Amazon.com, "Amazon Silk," 2011, [Online]. Available: http://www.amazon.com/silk Last Accessed: 2 November 2011.

[92] X. Jin and Y.-K. Kwok, "Cloud Assisted P2P Media Streaming for Bandwidth Constrained Mobile Subscribers," in *2010 IEEE 16th International Conference on Parallel and Distributed Systems.* IEEE, Dec. 2010, pp. 800–805.

[93] D. C. Barboza, H. L. Junior, E. W. G. Clua, and V. E. Rebello, "A Simple Architecture for Digital Games on Demand Using Low Performance Resources under a Cloud Computing Paradigm," in *2010 Brazilian Symposium on Games and Digital Entertainment.* IEEE, Nov. 2010, pp. 33–39.

[94] I. Nave, H. David, A. Shani, Y. Tzruya, A. Laikari, P. Eisert, and P. Fechteler, "Games@large graphics streaming architecture," in *2008 IEEE International Symposium on Consumer Electronics.* IEEE, Apr. 2008, pp. 1–4.

[95] OnLive Ltd., "OnLive," 2011, [Online]. Available: http://www.onlive.com Last Accessed: 26 October 2011.

[96] Gaikai Inc., "Gaikai Open Cloud Gaming Platform," 2011, [Online]. Available: http://www.gaikai.com Last Accessed: 26 October 2011.

[97] X. Luo, "From Augmented Reality to Augmented Computing: A Look at Cloud-Mobile Convergence," in *2009 International Symposium on Ubiquitous Virtual Reality.* IEEE, Jul. 2009, pp. 29–32.

[98] S. Wang and S. Dey, "Rendering Adaptation to Address Communication and Computation Constraints in Cloud Mobile Gaming," in *2010 IEEE Global Telecommunications Conference GLOBECOM 2010.* IEEE, Dec. 2010, pp. 1–6.

[99] M. Jarschel, D. Schlosser, S. Scheuring, and T. Hossfeld, "An Evaluation of QoE in Cloud Gaming Based on Subjective Tests," in *2011 Fifth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing.* IEEE, Jun. 2011, pp. 330–335.

[100] M. Aoyama, "Agile Software Process and its experience," in *Proceedings of the 20th International Conference on Software Engineering.* IEEE Comput. Soc, 1998, pp. 3–12.

[101] G. Huerta-Canepa and D. Lee, "A virtual cloud computing provider for mobile devices," in *Proceedings of the 1st ACM Workshop on Mobile Cloud Computing & Services Social Networks and Beyond - MCS '10*, ser. MCS '10. New York, New York, USA: ACM Press, 2010, pp. 1–5.

[102] D. Chu and M. Humphrey, "Mobile OGSI.NET: Grid Computing on Mobile Devices," in *Fifth IEEE/ACM International Workshop on Grid Computing.* IEEE, 2004, pp. 182–191.

[103] M. Black and W. Edgar, "Exploring mobile devices as Grid resources: Using an x86 virtual machine to run BOINC on an iPhone," in *2009 10th IEEE/ACM International Conference on Grid Computing.* IEEE, Oct. 2009, pp. 9–16.

[104] D. G. Rosado, E. Fernandez-Medina, J. L, and M. Piattini, "PSecGCM: Process for the Development of Secure Grid Computing based Systems with Mobile Devices," in *2008 Third International Conference on Availability, Reliability and Security.* IEEE, Mar. 2008, pp. 136–143.

[105] J. Park, H. Yu, K. Chung, and E. Lee, "Markov Chain Based Monitoring Service for Fault Tolerance in Mobile Cloud Computing," in *2011 IEEE Workshops of International Conference on Advanced Information Networking and Applications.* IEEE, Mar. 2011, pp. 520–525.

[106] Google Inc., "Google Goggles," 2011, [Online]. Available: http://www.google.com/mobile/goggles
Last Accessed: 8 November 2011.

[107] X. Gu, K. Nahrstedt, A. Messer, I. Greenberg, and D. Milojicic, "Adaptive offloading inference for delivering applications in pervasive

computing environments," in *Proceedings of the First IEEE International Conference on Pervasive Computing and Communications, 2003. (PerCom 2003).* IEEE Comput. Soc, 2003, pp. 107–114.

[108] R. K. Balan, M. Satyanarayanan, S. Y. Park, and T. Okoshi, "Tactics-based remote execution for mobile computing," in *Proceedings of the 1st international conference on Mobile systems, applications and services - MobiSys '03.* New York, New York, USA: ACM Press, May 2003, pp. 273–286.

[109] R. K. Balan, D. Gergle, M. Satyanarayanan, and J. Herbsleb, "Simplifying cyber foraging for mobile devices," in *Proceedings of the 5th international conference on Mobile systems, applications and services - MobiSys '07.* New York, New York, USA: ACM Press, Jun. 2007, p. 272.

[110] C. Xian, Y.-H. Lu, and Z. Li, "Adaptive computation offloading for energy conservation on battery-powered systems," in *2007 International Conference on Parallel and Distributed Systems.* IEEE, Dec. 2007, pp. 1–8.

[111] Amazon Web Services, "Amazon Web Services Software Development Kit for Android," 2011, [Online]. Available: http://aws.amazon.com/sdkforandroid/ Last Accessed: 15 November 2011.

[112] H. Mao, N. Xiao, W. Shi, and Y. Lu, "Wukong: Toward a Cloud-Oriented File Service for Mobile Devices," in *2010 IEEE International Conference on Services Computing.* IEEE, Jul. 2010, pp. 498–505.

[113] Y. J. Nam, Y. K. Park, J. T. Lee, and F. Ishengoma, "Cost-Aware Virtual USB Drive: Providing Cost-Effective Block I/O Management Commercial Cloud Storage for Mobile Devices," in *2010 13th IEEE*

International Conference on Computational Science and Engineering. IEEE, Dec. 2010, pp. 427–432.

[114] Dropbox Inc., "Dropbox - Mobile," 2011, [Online]. Available: https://www.dropbox.com/anywhere
Last Accessed: 11 November 2011.

[115] Box Inc., "Box for Mobile," 2011, [Online]. Available: http://www.box.net/features/mobile/
Last Accessed: 11 November 2011.

[116] M. Satyanarayanan, "Pervasive computing: vision and challenges," *IEEE Personal Communications*, vol. 8, no. 4, pp. 10–17, 2001.

[117] S. Goyal and J. Carter, "A Lightweight Secure Cyber Foraging Infrastructure for Resource-Constrained Devices," in *Sixth IEEE Workshop on Mobile Computing Systems and Applications.* IEEE, 2004, pp. 186–195.

[118] Amazon Web Services, "Amazon Web Services Direct Connect," 2011, [Online]. Available: http://aws.amazon.com/directconnect/
Last Accessed: 18 December 2011.

[119] GSM Association, "Digital Dividend for Mobile: Bringing Broadband to All," Tech. Rep., 2009, [Online]. Available: http://www.gsmworld.com/documents/dd_spectrumoverview.pdf
Last Accessed: 4 December 2011.

[120] Ericsson, "More Than 50 Billion Connected Devices," Tech. Rep., 2011, [Online]. Available: http://www.ericsson.com/res/docs/whitepapers/wp-50-billions.pdf
Last Accessed: 4 December 2011.

[121] M. Satyanarayanan, "Mobile Computing: The Next Decade," in *Proceedings of the 1st ACM Workshop on Mobile Cloud Computing & Services Social Networks and Beyond - MCS '10.* New York, New York, USA: ACM Press, Jun. 2010, pp. 1–6.

Appendix A

Data Pack

This minor dissertation is accompanied by a data pack containing relevant information in digital format.

The purpose of this additional information is to supplement this minor dissertation with material created in the course of this research endeavour.

This data pack is available on the optical disc which should accompany hard-copy editions of this work or as an archive file accompanying digital editions.

The text file named *"Readme.txt"* in the root directory this data pack contains detailed instructions regarding the following:

- The location of content within the file structure.
- Use of the supplied hash values for verifying the integrity of this content.
- Possible mechanisms for accessing and using this content.

The following additional information and content is contained within this data pack:

- **LaTeX Source:** All LaTeXsource files which constitute this minor dissertation are included in the data pack. These can be used for reference or to recompile the main document as required.

- **Image Files:** All images used in this document are included in the data pack. The majority of these images have been created by the author and are thus provided under the same licence terms as this document. Where images from external sources have been included, the relevant licence terms are applicable.

- **Conference Paper:** A conference paper on the subject of this research was accepted for presentation at the Southern Africa Telecommunications, Applications and Networks (SATNAC) 2011 conference which was held in East London, South Africa in September 2011. The original conference paper as well as the LaTeX source files are included in this data pack. As a published document, this conference paper is subject to the licence terms of the SATNAC 2011 proceedings.

- **Conference Presentation:** The presentation delivered at SATNAC 2011 on the topic of this research is included in this data pack. This presentation is subject to the same licence terms as this minor dissertation.

- **Conference Report:** A short report on the author's participation at SATNAC 2011 has been compiled for funding purposes. This document has not previously been published and is available in the data pack

- **Research Overviews:** In the course of this research endeavour, two overview documents have been compiled. The purpose of these was to communicate the important aspects of this research to external audiences. Both technical and non-technical overviews are included in this data pack. These documents have not previously been published.